

# Clustering

Jaime Arguello

INLS 613: Text Data Mining

[jarguell@email.unc.edu](mailto:jarguell@email.unc.edu)

October 9, 2013

(some material taken or adapted from slides by Hinrich Schutze)

# Clustering

## objective

- Grouping documents or instances into subsets or clusters
- Documents within a the same cluster should be similar
- Documents from different clusters should be dissimilar
- A common form of **unsupervised learning**
- Unsupervised = no human-produced labels
- The goal is to discover structure from the data

# Clustering vs. Classification

- Classification:
  - ▶ the input to the system is a set of labeled data
  - ▶ the algorithm learns a model for predicting the label on new examples
- Clustering:
  - ▶ the input the system is a set of unlabeled data
  - ▶ the algorithm infers the labels from the data and assigns a label to each input instance

# Clustering applications

- **Search engine results clustering:** grouping search engine results by topic
  - ▶ the user can identify the relevant clusters and ignore the non-relevant ones
- **Collection clustering:** grouping documents by topic to support navigation and exploration
- **Data analytics:** grouping instances to identify popular trends (big clusters) and outliers (small clusters)

# Clustering Applications

## search engine results clustering

The screenshot shows the Yippy search engine interface. At the top, there's a navigation bar with links for 'web', 'news', 'images', 'maps', 'blogs', 'wikipedia', 'jobs', and 'more'. The search bar contains the word 'jaguar', and there are buttons for 'Search' and 'advanced preferences'. Below the search bar, there's a sidebar on the left with tabs for 'clouds', 'sources', 'sites', and 'time'. The 'clouds' tab is active, showing a list of clusters for the query 'jaguar'. The clusters include 'Jaguar Cars (24)', 'International (13)', 'Pictures (27)', 'Panthera, Onca (17)', 'Parts (23)', 'Dealership, Sells and services Jaguar (22)', 'Luxury, Car (15)', 'Club, Events (17)', 'Jaguar Land Rover (9)', and 'Reviews, Prices (8)'. There's a 'remix' button and a 'find in clouds' search bar. The main content area displays the top 170 results of at least 202,000,000 retrieved for the query 'jaguar'. It includes a definition of 'jaguar' as a noun, a link to 'See more from Encyclopedia', and several sponsored results from Jaguar's official site and other related websites. At the bottom, there's a font size selector and a link to 'Jaguar International - Market selector page'.

web news images maps blogs wikipedia jobs more »

jaguar Search advanced preferences

clouds sources sites time remix

All Results (176)

- + Jaguar Cars (24)
- + International (13)
- + Pictures (27)
- + Panthera, Onca (17)
- + Parts (23)
- + Dealership, Sells and services Jaguar (22)
- + Luxury, Car (15)
- + Club, Events (17)
- + Jaguar Land Rover (9)
- + Reviews, Prices (8)

more | all clouds

find in clouds: Find

Font size: A A A A

Top 170 results of at least 202,000,000 retrieved for the query **jaguar** (definition) (details)

jaguar

- noun - jaguar, panther, Panthera onca, Felis onca -- (a large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis)

[See more from Encyclopedia »](#)

**Jaguars**

JAGUARS JAGUARS . The jaguar (Panthera onca ) is the largest native American cat, and for over three thousand years it has been one of Central and South America's most important symbolic animals. Sometimes associated with the puma (Felis concolor ) and ocelot (Felis pardalis ), the jaguar was a

**Jaguar® Official Site**

Experience How Luxury Feels When You Build & Price A Jaguar.  
[www.jaguarusa.com](http://www.jaguarusa.com)

**Jaguar**

Feel how alive luxury can be - Drive our breathtaking 2012 models  
[www.flowjaguargreensboro.com](http://www.flowjaguargreensboro.com)

**Consider a Mercedes-Benz®**

Discover the Safety & Performance Innovations That Set the Benchmark.  
[www.mbusa.com/Raleigh-Durham](http://www.mbusa.com/Raleigh-Durham)

**Jaguar International - Market selector page**



Jaguar Cars Limited: Registered Office: Abbey Road, Whitley, Coventry CV3 4LF Registered in England  
No: 1672070. You need Flash Player 9  
[www.jaguar.com/gl/en/marketsel](http://www.jaguar.com/gl/en/marketsel) - [cache] - Additional Sources, Yippy Sources

Sponsored Results

Search Results

# Clustering Applications

## collection clustering



**News**

U.S. edition ▾Modern ▾

**Top Stories**

Mitt Romney

Chromebook

Washington Redskins

Earthquake

Fidel Castro

Cleveland Browns

George McGovern

Toronto Blue Jays

Brad Pitt

Jay-Z

North Carolina

World

U.S.

Business

Elections

Technology

Entertainment


Sports

Science

Health

Spotlight

**Top Stories**



The Guardian

See realtime coverage

### Police chief: Wisconsin spa shooting suspect died of self-inflicted wound

Chicago Tribune - 23 minutes ago

A man police suspected of killing three and wounding four by opening fire at a tranquil day spa was found dead Sunday afternoon following a six-hour manhunt that locked down a shopping center, country club and hospital in suburban Milwaukee.

[Suspect in Wisconsin spa shooting found dead](#) Fox News


[Three Killed in Shooting at Spa in Wisconsin](#) New York Times

Related [Brown Deer, Wisconsin »](#)

Highly Cited: [3 killed, 4 injured in rampage at Azana Spa in Brookfield](#) Milwaukee Journal Sentinel

In Depth: [Three killed in shooting at Milwaukee-area salon; suspect found dead at scene](#) NBCNews.com


Wikipedia: [2012 Azana Spa shootings](#)

 ABC News


4 hours ago - Google+

[Breaking: The area is on lockdown as authorities in Brookfield work to secure the scene, which is across the street from a shopping mall in Wisconsin.](#)


[At Least 7 Injured in Spa Shooting](#)




The Associated P...




The Associated P...




YouTube




CNN (blog)




CBS News



Christian S...



Newsday



Wall Street ...

### Romney, Obama in Dead Heat

Wall Street Journal - 22 minutes ago

By NEIL KING JR. Mitt Romney has strengthened his image as the candidate best able to boost the economy and has fought President Barack Obama to a near-draw on who can best serve as commander in chief, helping turn the 2012 election into a tie ...



6

Wednesday, October 9, 13



# Clustering Applications

## collection clustering



**News**


U.S. edition ▾Modern ▾


**Top Stories**

- Mitt Romney
- Chromebook**
- Washington Redskins
- Earthquake
- Muammar al-Gaddafi
- Fidel Castro
- Cleveland Browns
- George McGovern
- Brad Pitt
- Toronto Blue Jays

**News near you**

- World
- U.S.
- Business
- Elections
- Technology
- Entertainment
- Sports
- Science
- Health
- Spotlight



**Chromebook**



ZDNet

[See realtime coverage](#)

### Review: The ARM-powered Samsung Chromebook








ZDNet 19 minutes ago |  Written by [Steven Vaughan-Nichols](#) 

I was already a big Chromebook fan before I got my hands on Samsung's just-released ARM-powered Chromebook. Now, after a weekend with it and with its amazing price of \$249 I think it's going to find a few million more fans.


[New Samsung Chromebook By Google: ARM-Powered Ultrabook initial ...](#) [PC-Tablet - by Nitin Agarwal](#)

[New Google, Samsung Laptop Is Cheaper But Will That Be Enough To Sway ...](#)

Laptop Computer Planet Blog (blog)




ZDNet ZDNet Laptop Co... PolicyMic eWeek Sky News ... eGov monitor




SlashGear

### Samsung Chromebook (late-2012) Review


SlashGear Oct 20, 2012 |  Written by [Chris Burns](#)

This piece of Samsung hardware is the most basic Chromebook you can buy right this minute, but it's not the low-quality piece of hardware the price suggests.

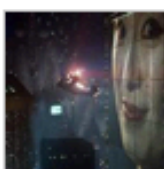


ZDNet

### New Samsung Chromebook and Samsung Series 5 550 head-to-head


ZDNet Oct 20, 2012 |  Written by [James Kendrick](#)

The new Samsung Chromebook is about the same thickness of the 550 (0.81 inches) but is much lighter at 2.43 pounds. It is very similar to the MacBook Air in both size and appearance.



ZDNet

### Computing's low-cost, Cloud-centric future is not Science Fiction

ZDNet 4 hours ago |  Written by [Jason Perlow](#)

This week, Google and Samsung released a new \$250 version of the Chromebook. There's not much new or even innovative about this particular device, particularly from a UX standpoint -- it's the same Chrome OS we've seen before, except that it now runs ...

# Clustering objective

- Grouping documents or instances into subsets or clusters
- Documents within a the same cluster should be similar
- Documents from different clusters should be dissimilar



# Clustering

## basics

- What does it mean for documents to be “similar” or “dissimilar”?

# Clustering

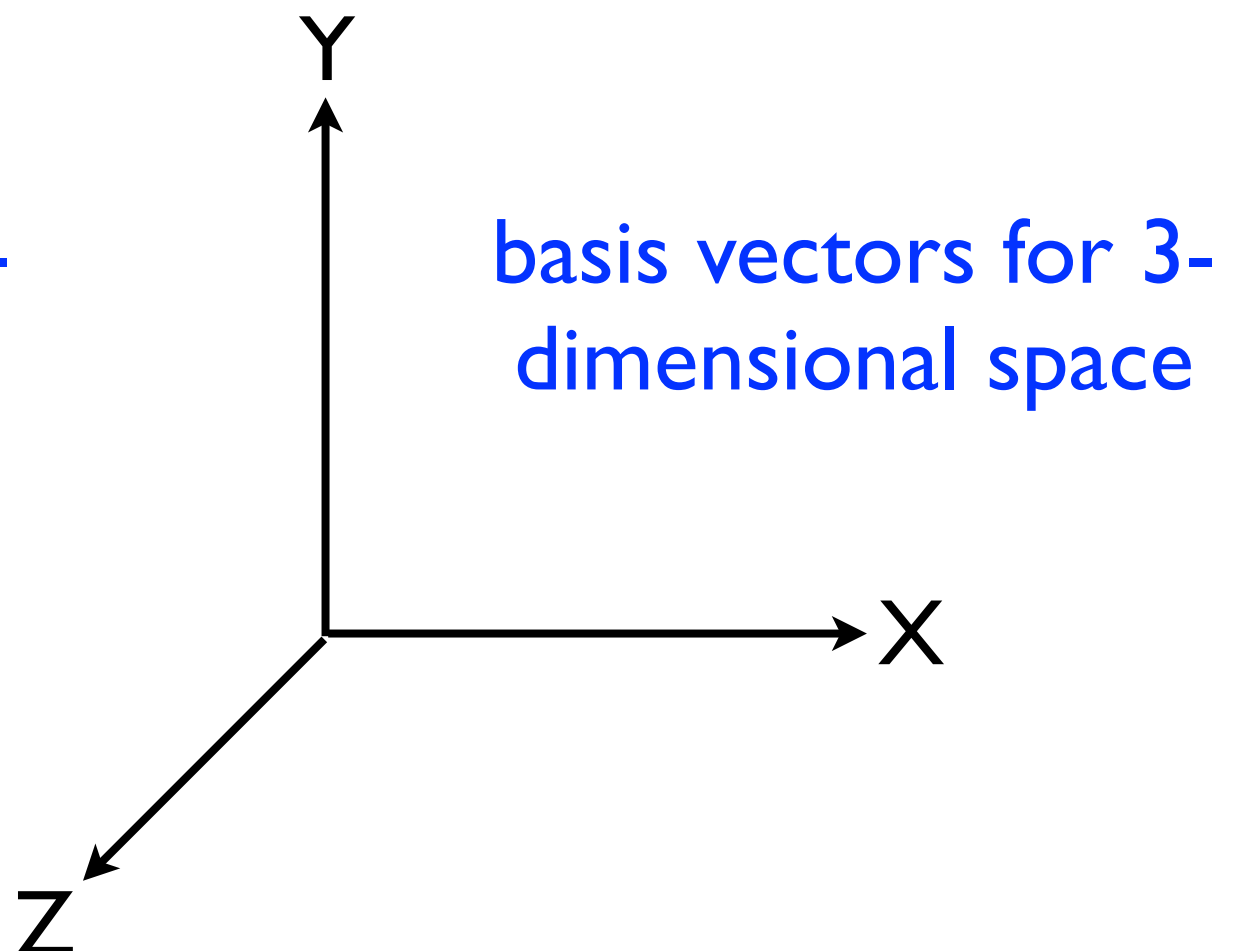
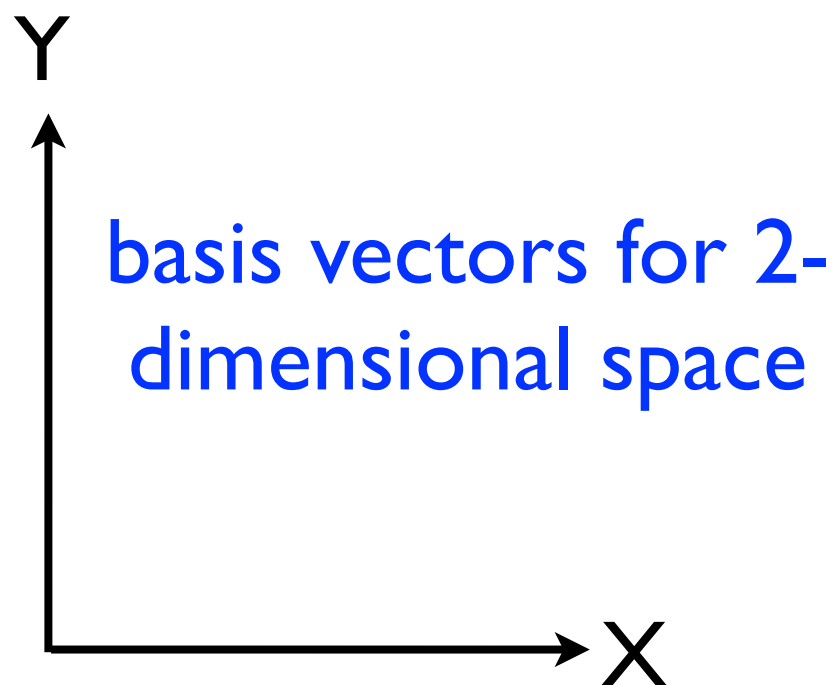
## basics

- What does it mean for documents to be similar or dissimilar?
- We need a computational way of modeling similarity
- **One solution:** model similarity using distance in a vector space representation of the collection or dataset
  - small distance = high similarity
  - long distance = low similarity

# Vector Space Representation

## review

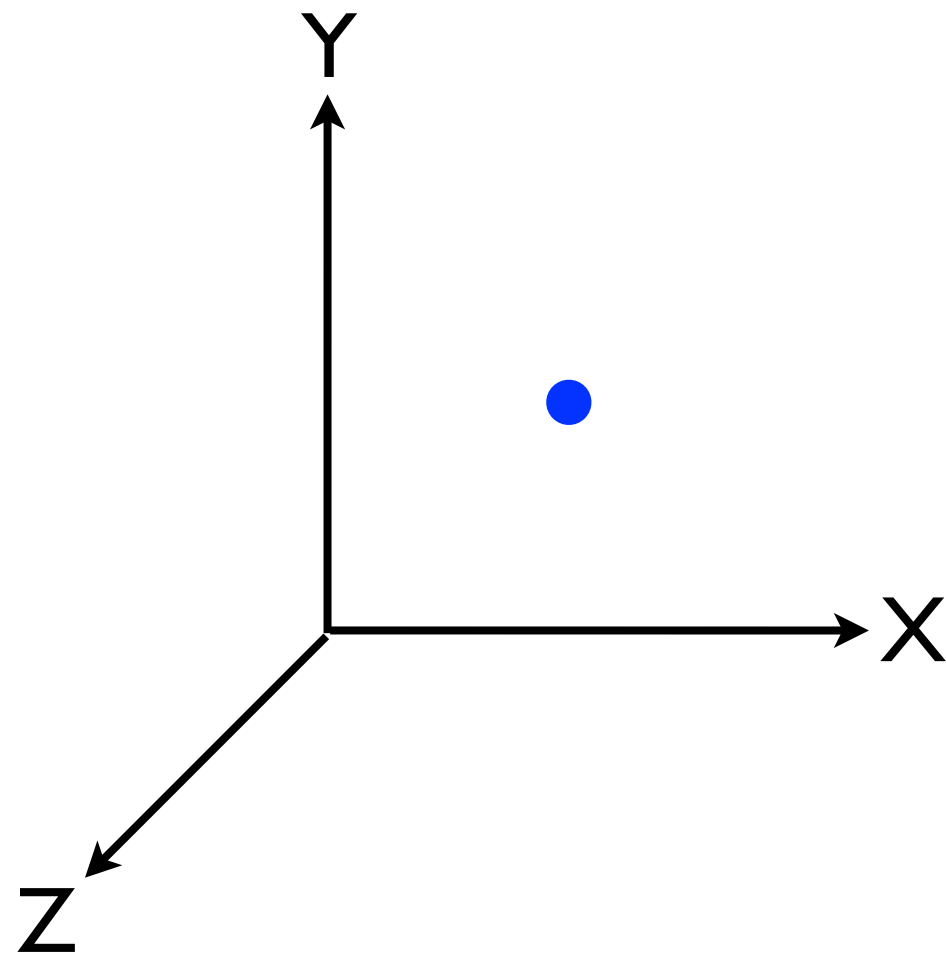
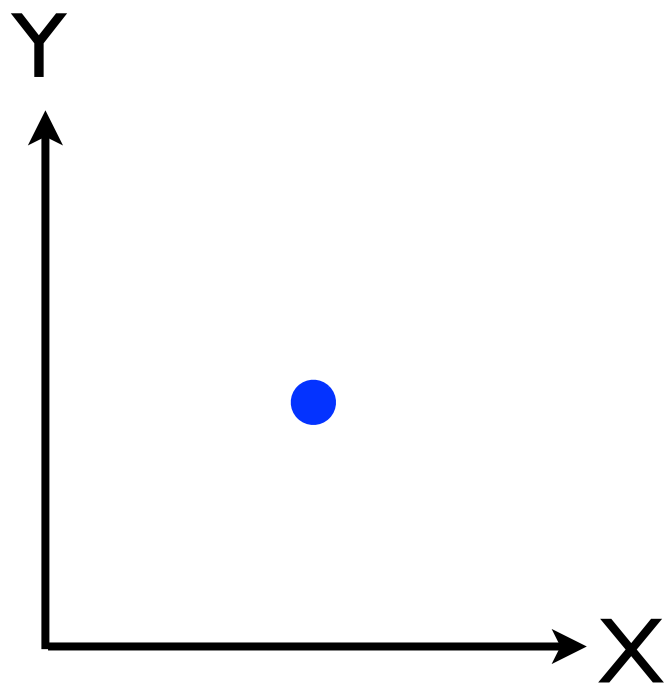
- A **vector space** is defined by a set of linearly independent basis vectors
- The **basis vectors** correspond to the dimensions or directions of the vector space



# Vector Space Representation

## review

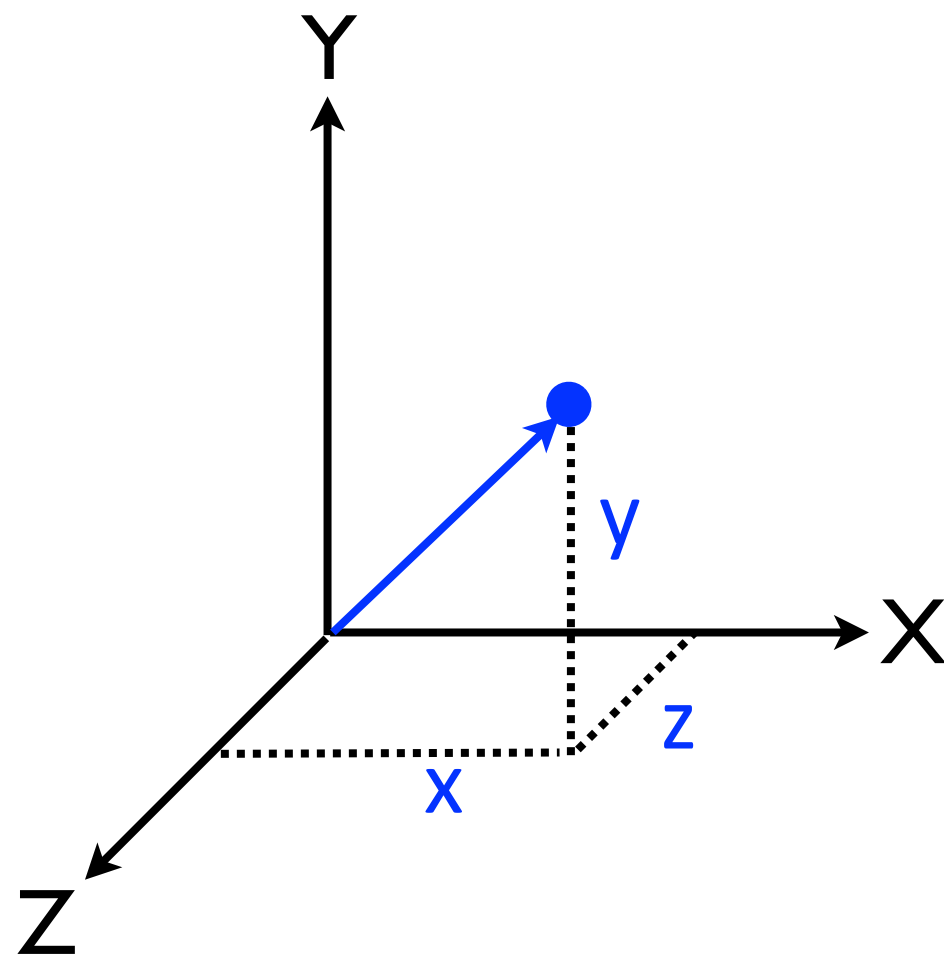
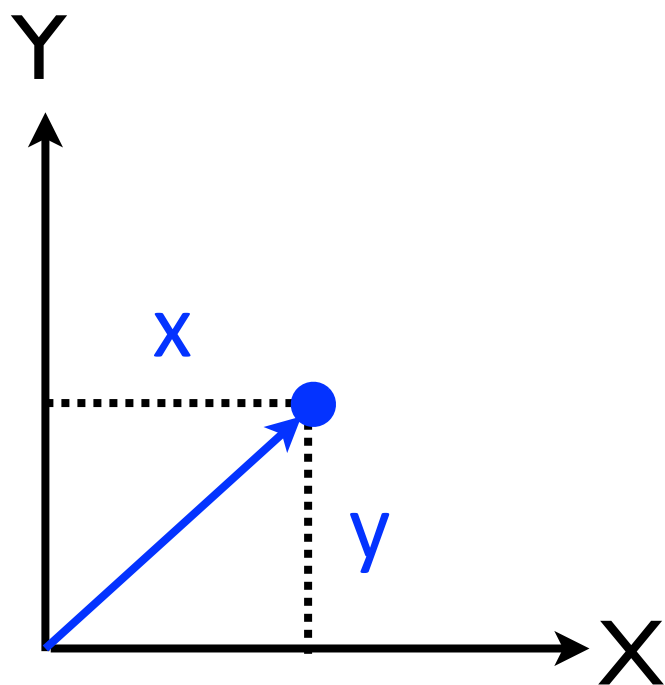
- A **vector** is a point in a vector space



# Vector Space Representation

## review

- A 2-dimensional vector can be written as  $[x,y]$
- A 3-dimensional vector can be written as  $[x,y,z]$



# Vector Space Representation

## review

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1



# Vector Space Representation

## review

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1

- We can represent this document as a vector in a 10-dimensional vector space

# Vector Space Representation

## review

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1

- This representation assumes binary term-weights.
- Are there other term-weighting schemes?



# TF, IDF, or TF.IDF?

adrian all already also an and apartment apollo as aspiring at  
balboa become better big boxer boxing but by can career champion  
chance creed current debt doesn't earns every exhibition extra far fight for gazzo gets girl  
go has he heavyweight her himself his in is it keep later life living loan lovers  
make man match meat men mickey named nobody of paulie pet philadelphia  
rocky set she shot small somebody someone still store struggling supplies surprised  
that the they think this through time title to trainer training up want when where  
who willing with woman won works



# TF, IDF, or TF.IDF?

ability adrain **adrian** already apartment **apollo** aspiring **balboa** become  
befriended befriends big **boxer** boxes **boxing** canvas champion chance checks  
chooses collecting collector **creed** current deadbeats debt debts distance doesn't downtown  
earns ease easily exhibition extra extremely factory fight forgot **gazzo** gear gotten  
**heavyweight** his is jergens later loan lot lovers managers match meat mickey named  
nobody odds packing paulie pennsylvania pet **philadelphia** pittance promoter  
publicity ready **rocky** sells set shark sharp shot shy somebody someone stallion store  
struggling stunt supplies supposed surprised thanksgiving think thrilled time title **touting** trainer training  
triumph up ve **viciousness** visits where who willing won works



# TF, IDF, or TF.IDF?

ability **adrain** adrian already apollo aspiring **balboa**  
beat **befriended** befriends better boxer **boxes** boxing  
**canvas** cash champion checks chooses **collecting**  
collector **creed** current **deadbeats** debt debts  
distance **doesn** downtown earns ease easily  
**exhibition** explains extra extremely factory far **forgot**  
**gazzo** gear giving gotten **heavyweight** idea interested  
italian **jergens** keep living loan lot lovers **managers** match meat  
mickey nobody odds **packing** paulie pennsylvania pet  
philadelphia **pittance** promoter prove **publicity**  
ready rocky sells shark sharp shop shy skills **somebody** spends  
**stallion** struggling **stunt** supplies supposed surprised  
thanksgiving think **thrilled** title **touting** trainer training  
triumph unknown **ve** **viciousness** visits want willing win  
won

# Vector Space Representation

## review

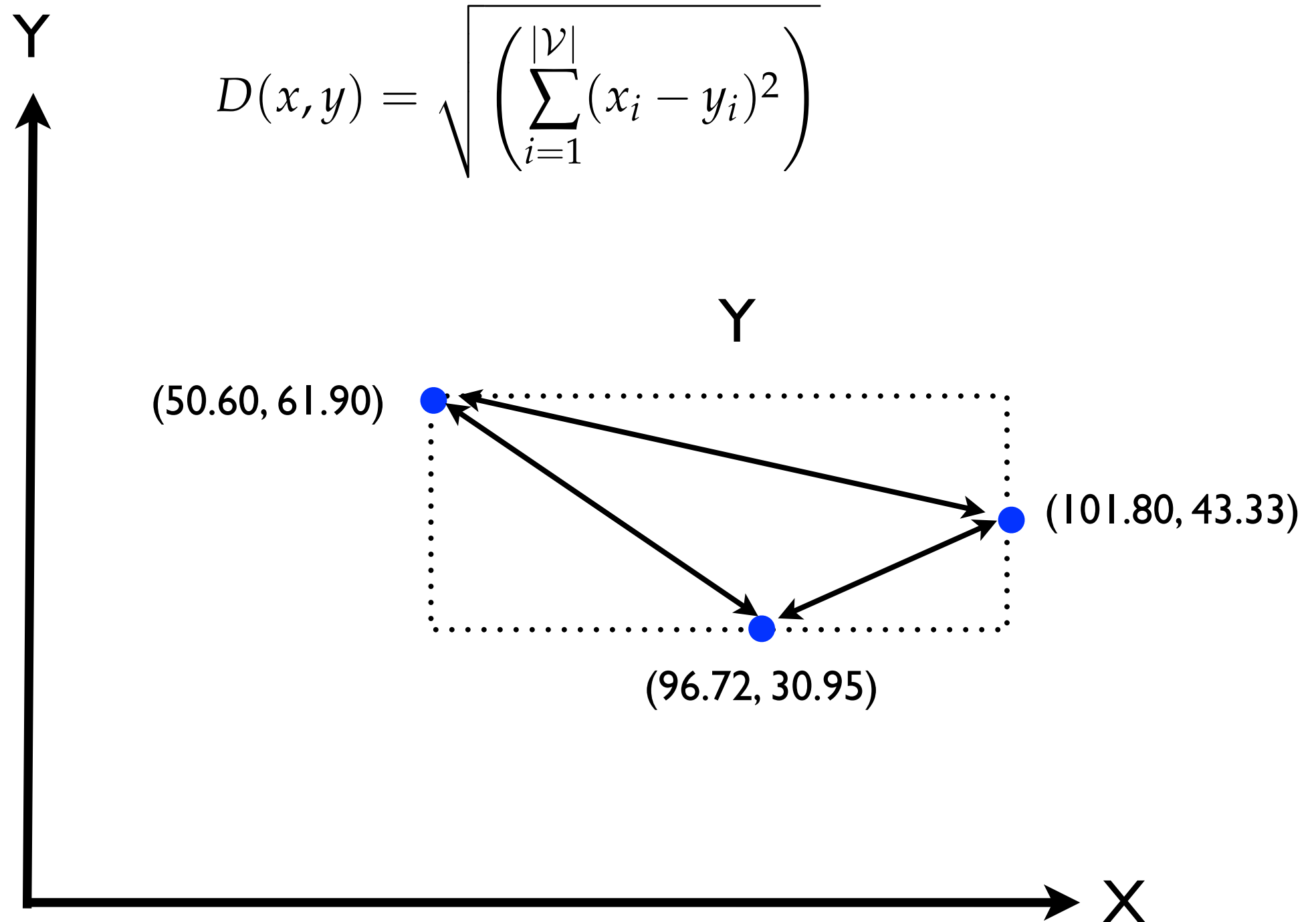
- Similarity = Euclidean Distance:

$$D(x, y) = \sqrt{\sum_{i=1}^{|\mathcal{V}|} (x_i - y_i)^2}$$



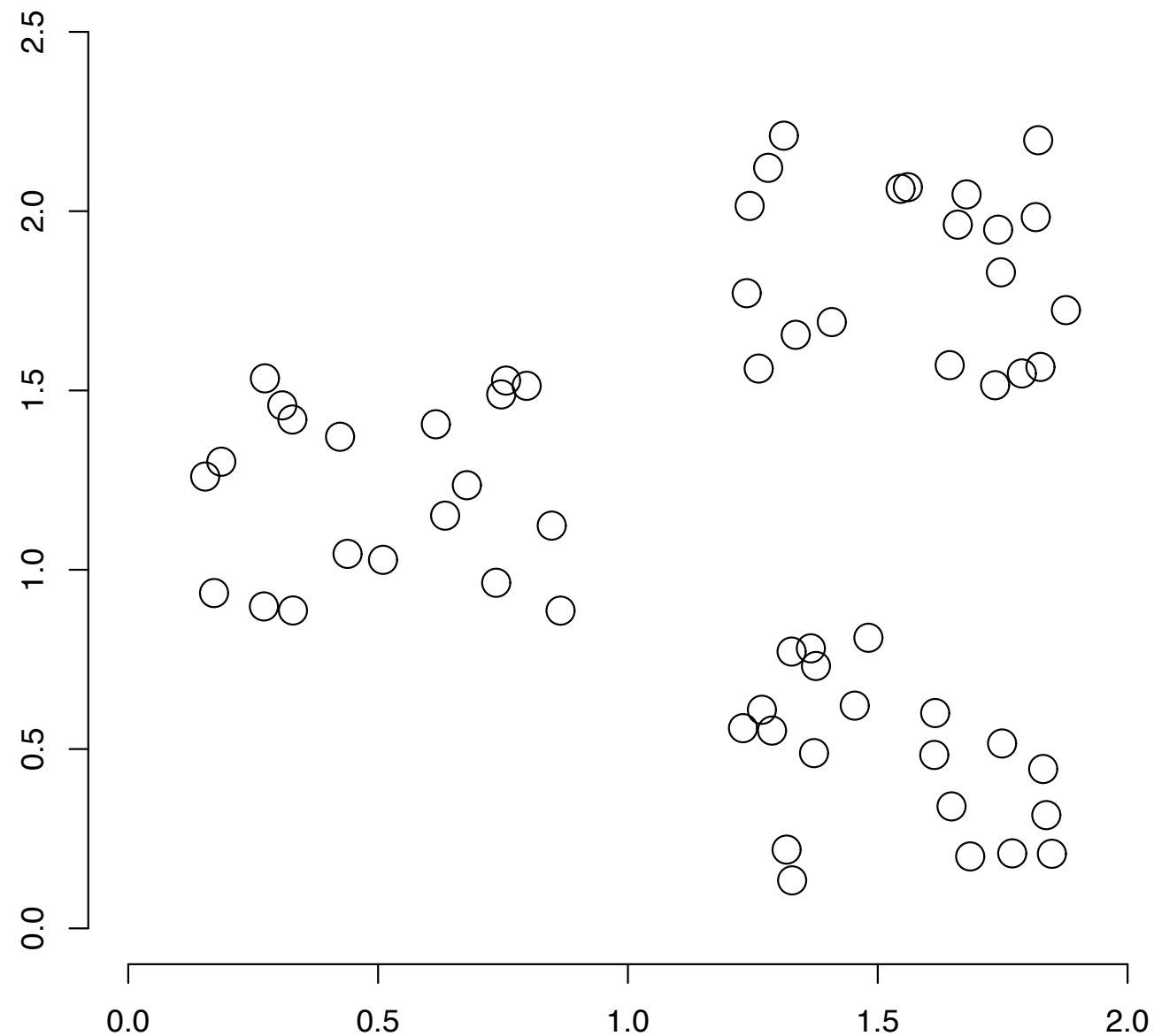
# Vector Space Representation

review



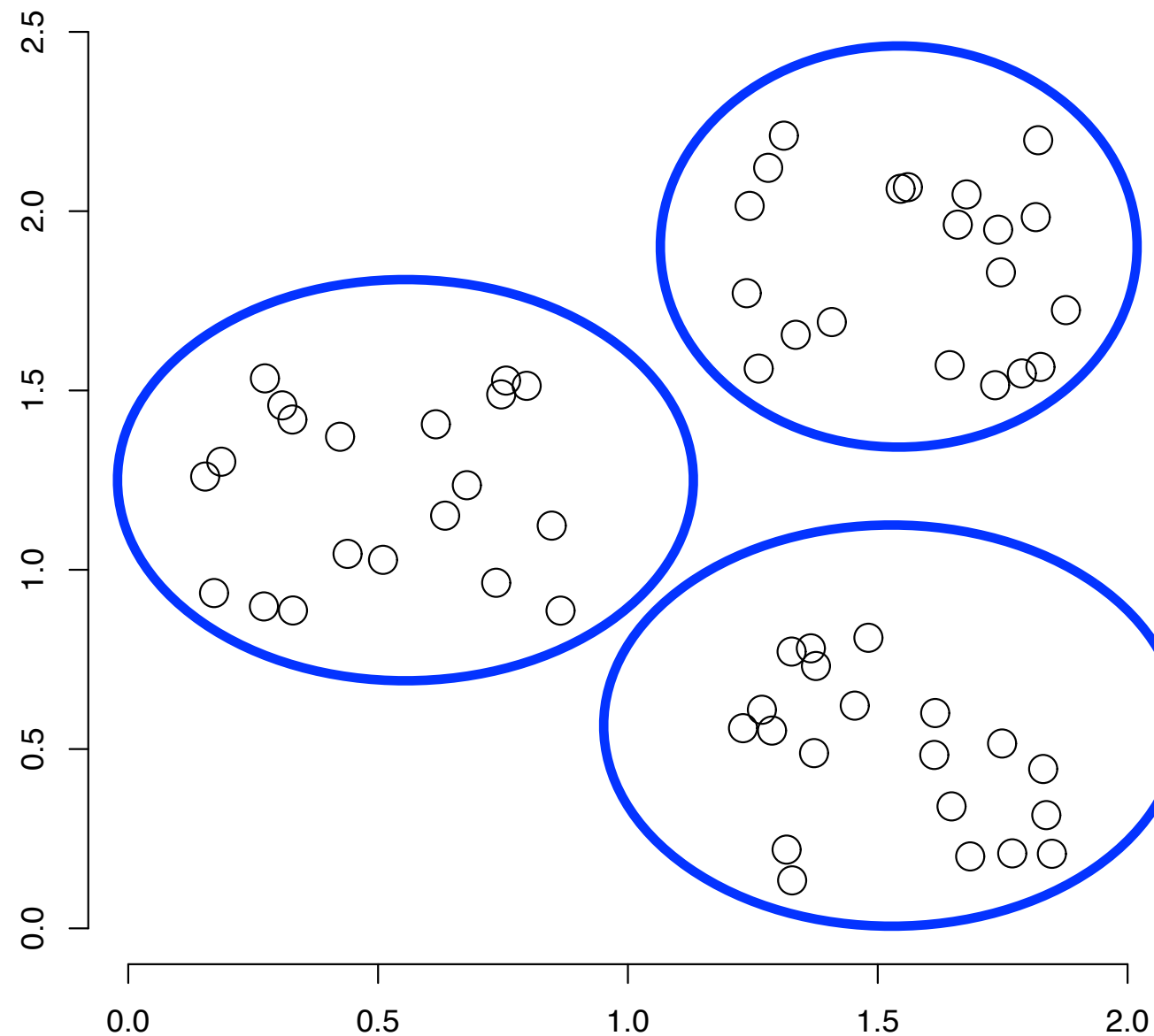
# Clustering

- What would we expect a clustering algorithm to do with this dataset?



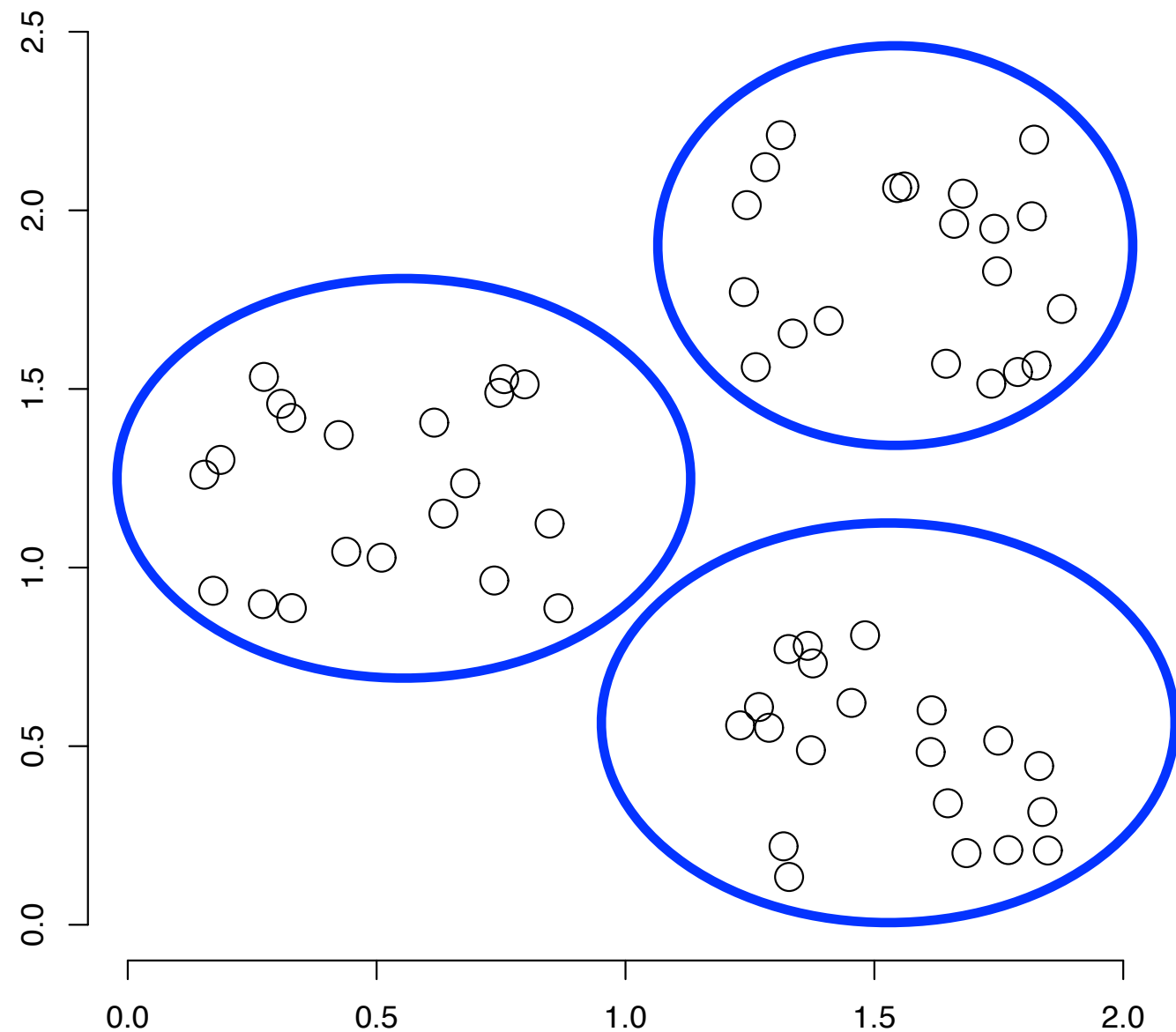
# Clustering

- What would we expect a clustering algorithm to do with this dataset?



# Clustering

- Propose an algorithm that might be able to do this!

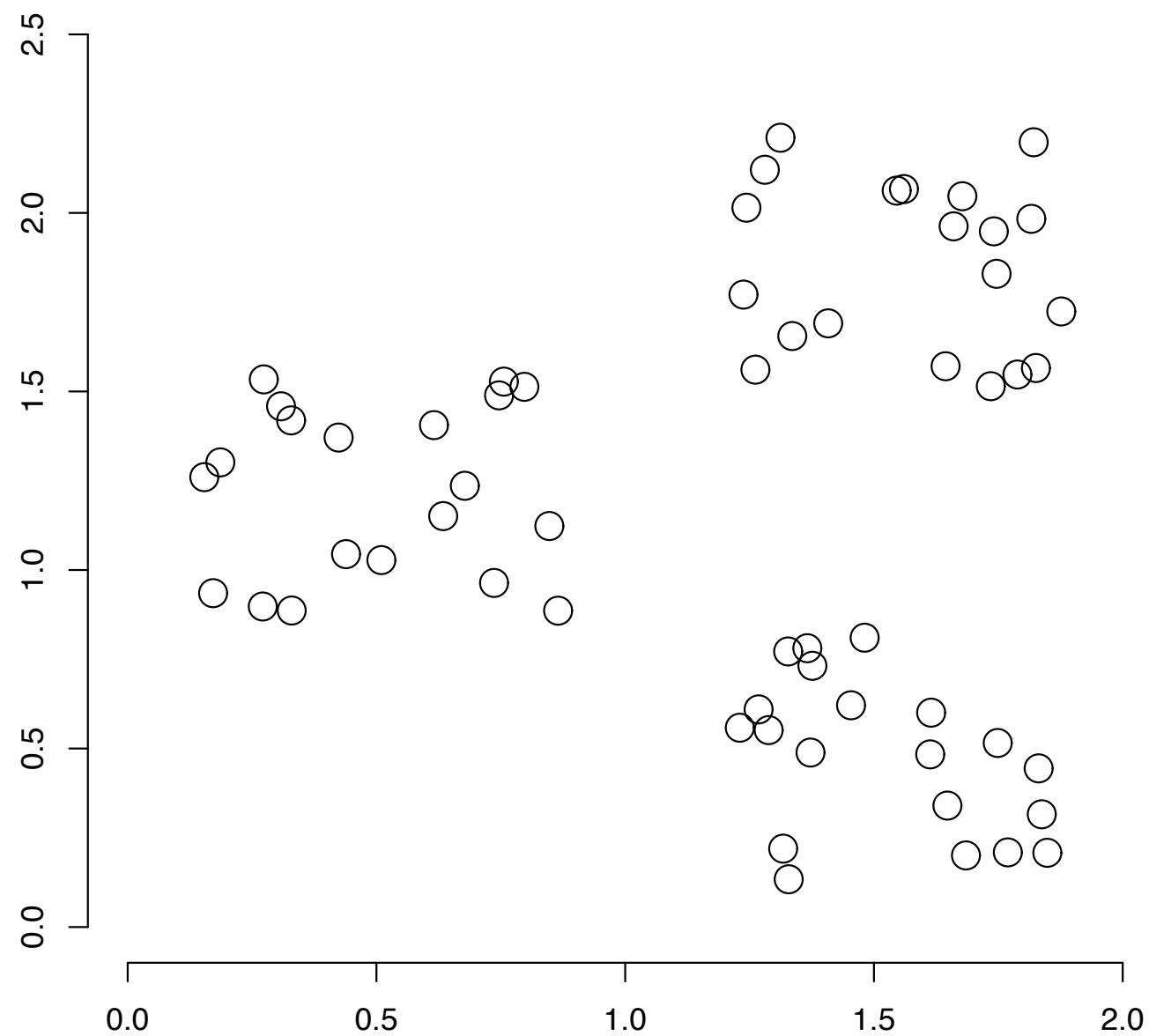


# Clustering

- **Input:** number of desired clusters  $K$
- **Output:** assignment of documents to  $K$  clusters
- **Algorithm:**
  - ▶ randomly select  $K$  documents (seeds)
  - ▶ assign each remaining document to its nearest seed

# Clustering

- Could this work?



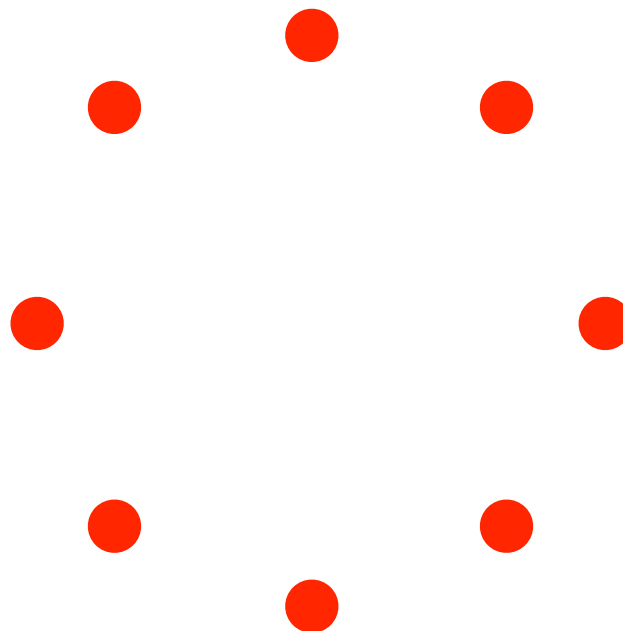


# K-Means Clustering

# K-means Clustering

## cluster centroid

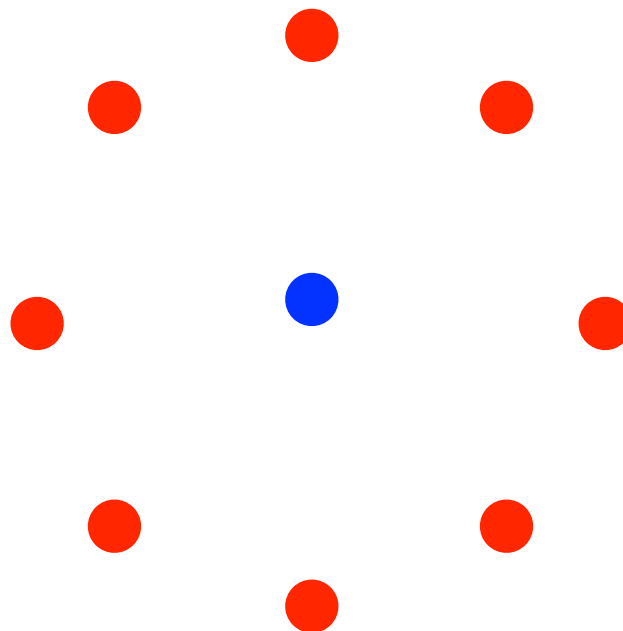
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

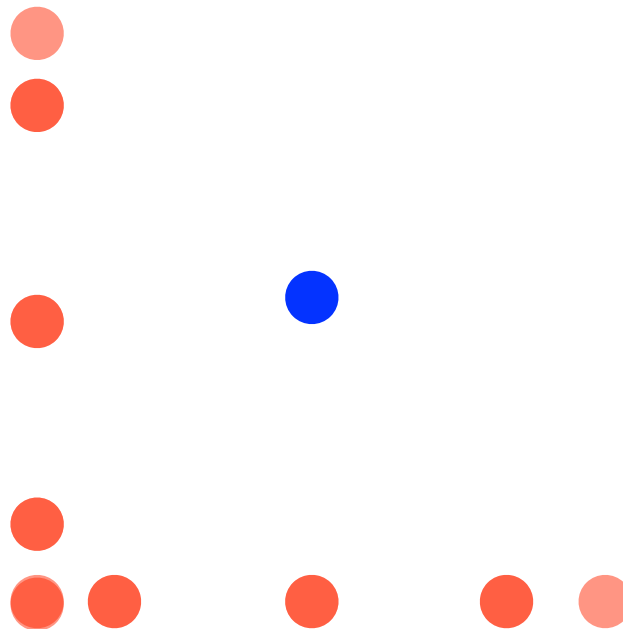
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

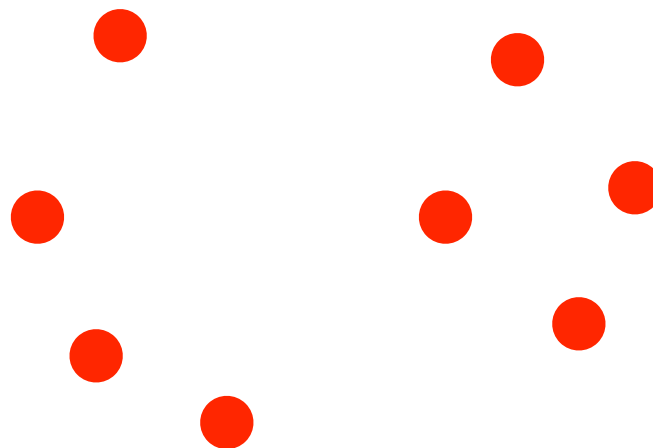
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”

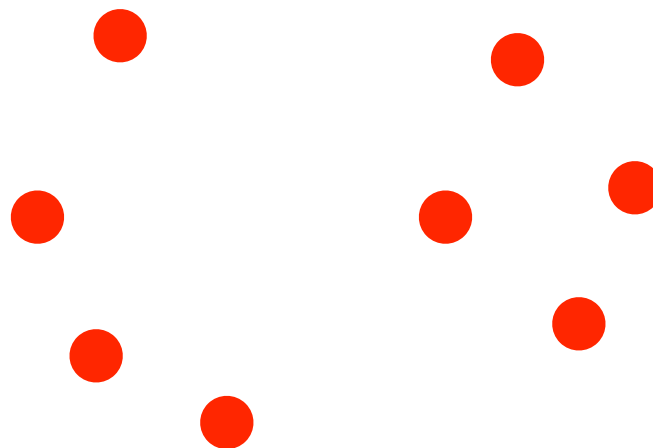




# K-means Clustering

## cluster centroid

- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

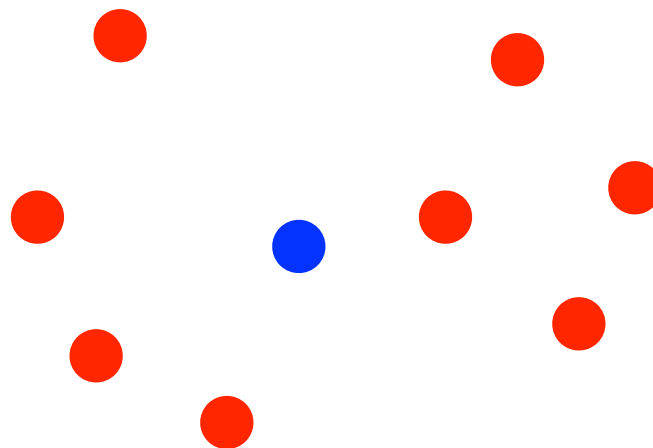
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

docs  
assigned to  
cluster 1

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
0	0	1	0	1	1	0	1	1	1
1	1	0	1	1	0	0	1	0	1

cluster 1  
centroid

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
?	?	?	?	?	?	?	?	?	?

# K-means Clustering

## cluster centroid

docs  
assigned to  
cluster 1

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
0	0	1	0	1	1	0	1	1	1
1	1	0	1	1	0	0	1	0	1

cluster 1  
centroid  
(average!)

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
0.33	0.50	0.50	0.50	1.00	0.33	0.33	0.83	0.50	0.50

# K-means Clustering

## cluster centroid

- For each dimension  $i$ , set:

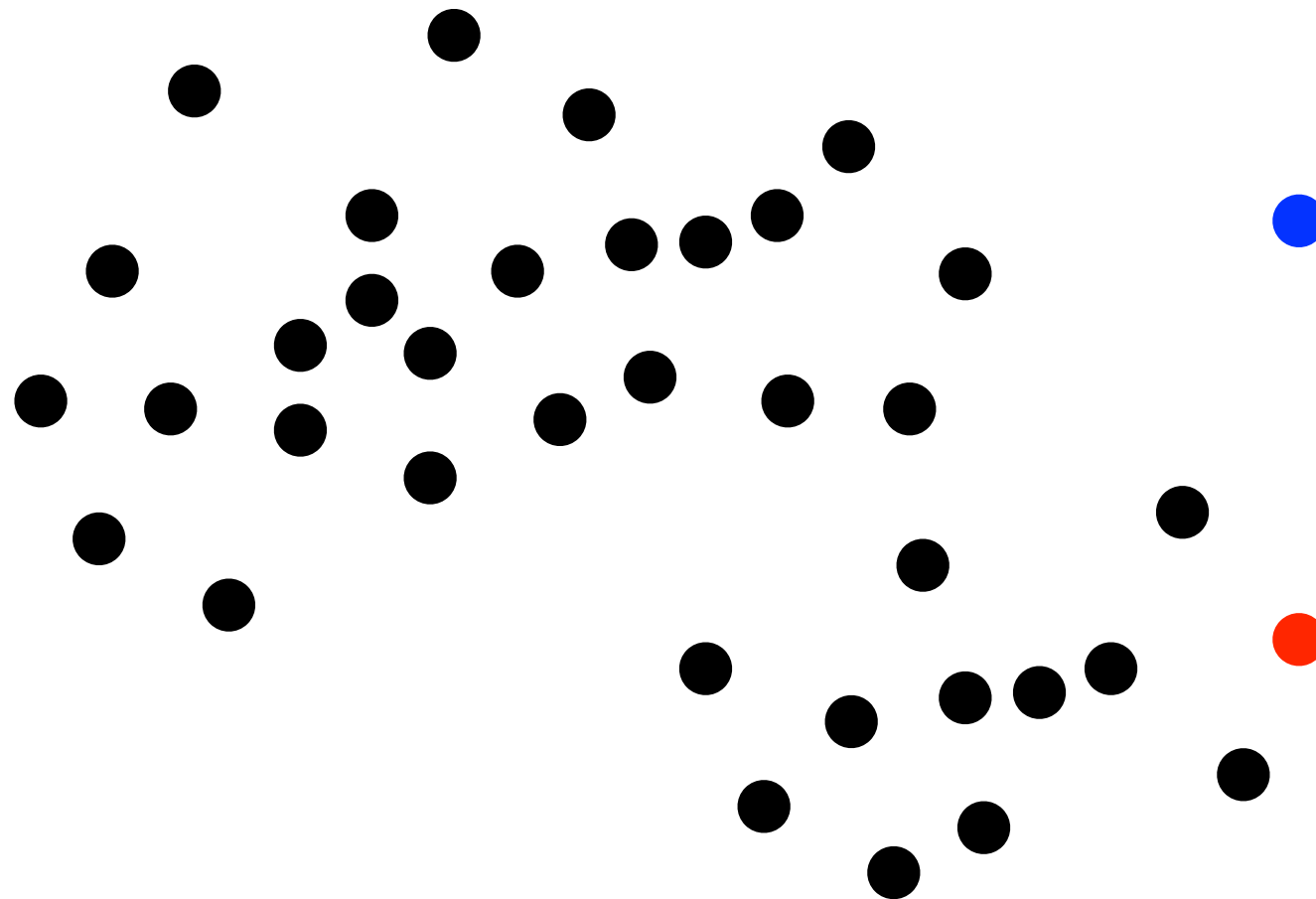
$$c_i = \frac{1}{|C|} \sum_{d \in C} d_i$$

# K-means Clustering

- **Input:** number of desired clusters  $K$
- **Output:** assignment of documents to  $K$  clusters
- **Algorithm:**
  - ▶ **Step 1:** randomly select  $K$  documents (seeds)
  - ▶ **Step 2:** assign each document to its nearest seed
  - ▶ **Step 3:** compute all  $K$  cluster centroids
  - ▶ **Step 4:** re-assign each document to its nearest centroid
  - ▶ **Step 5:** re-compute all  $K$  cluster centroids
  - ▶ **Step 6:** repeat steps 4 and 5 until terminating condition

# K-means Clustering

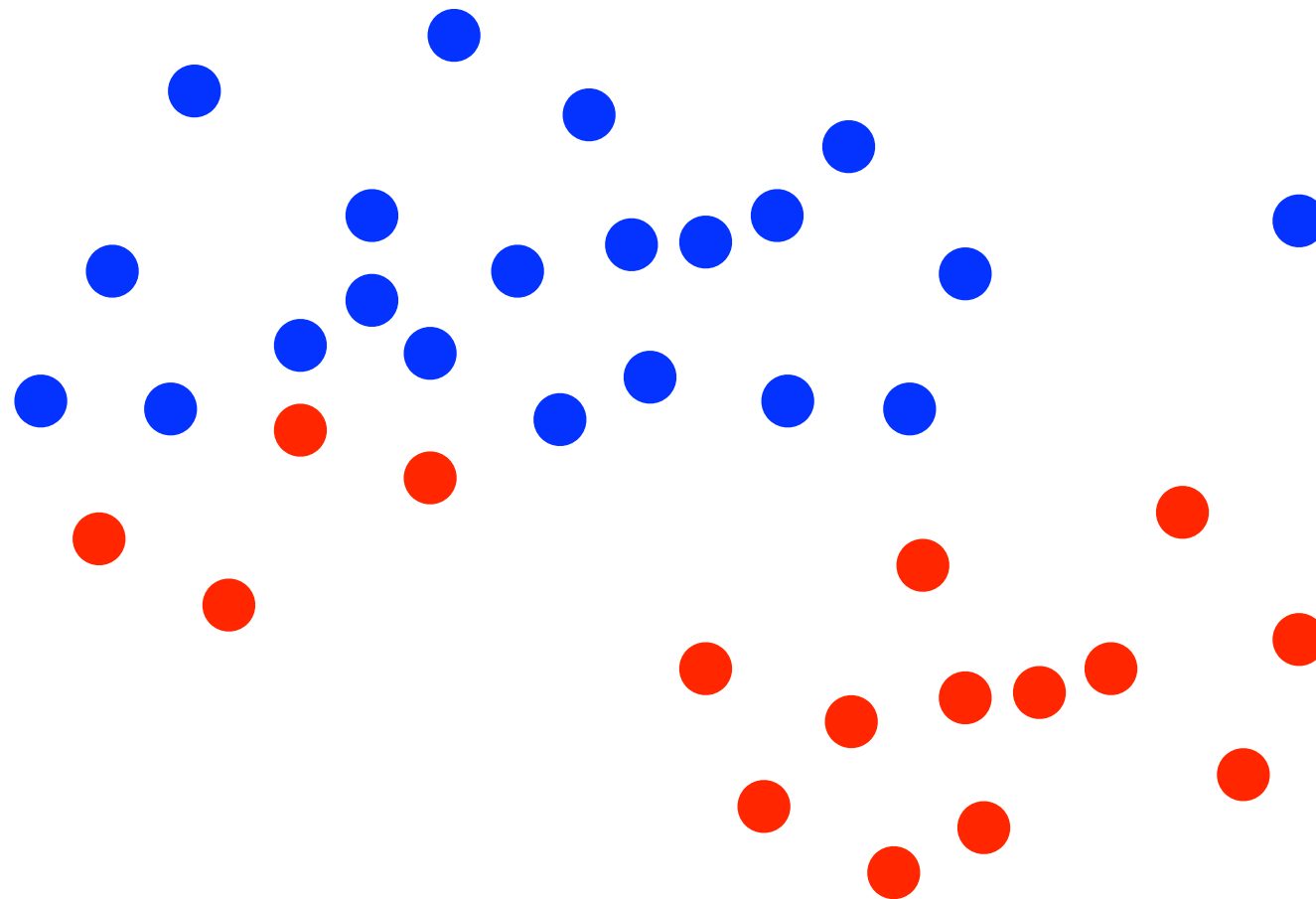
- Step 1: randomly select K documents (seeds)





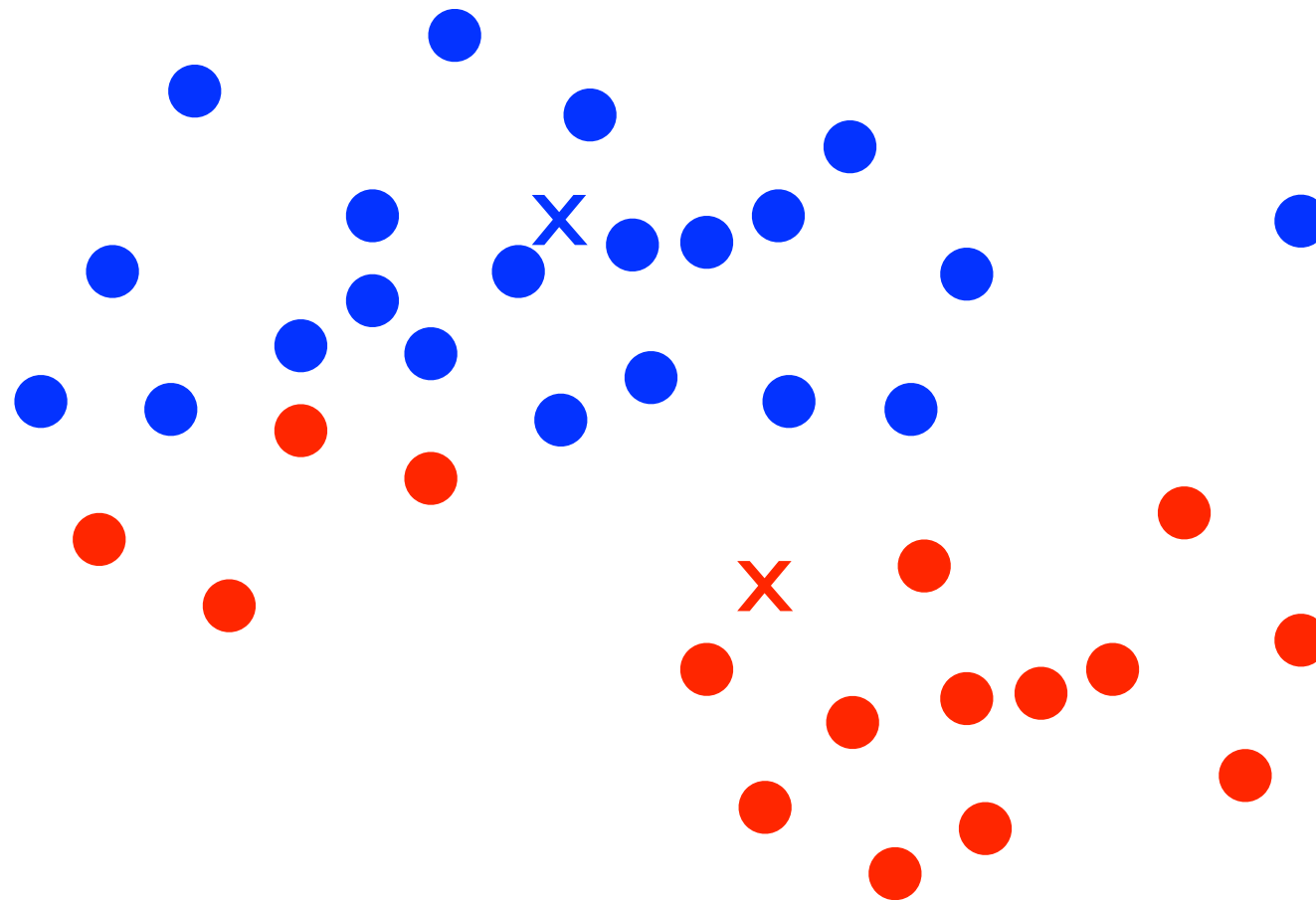
# K-means Clustering

- Step 2: assign each document to its nearest seed



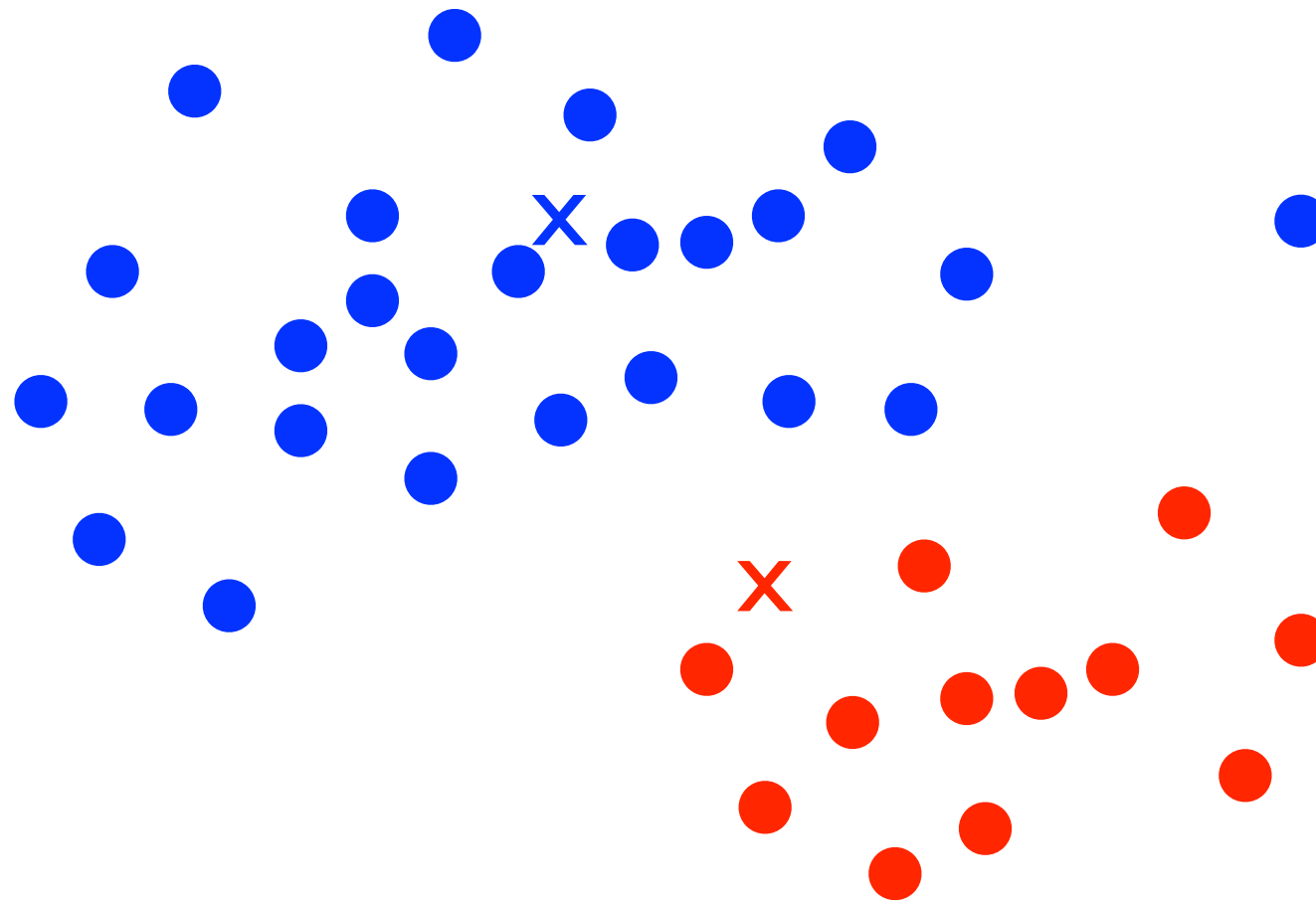
# K-means Clustering

- Step 3: compute all K cluster centroids



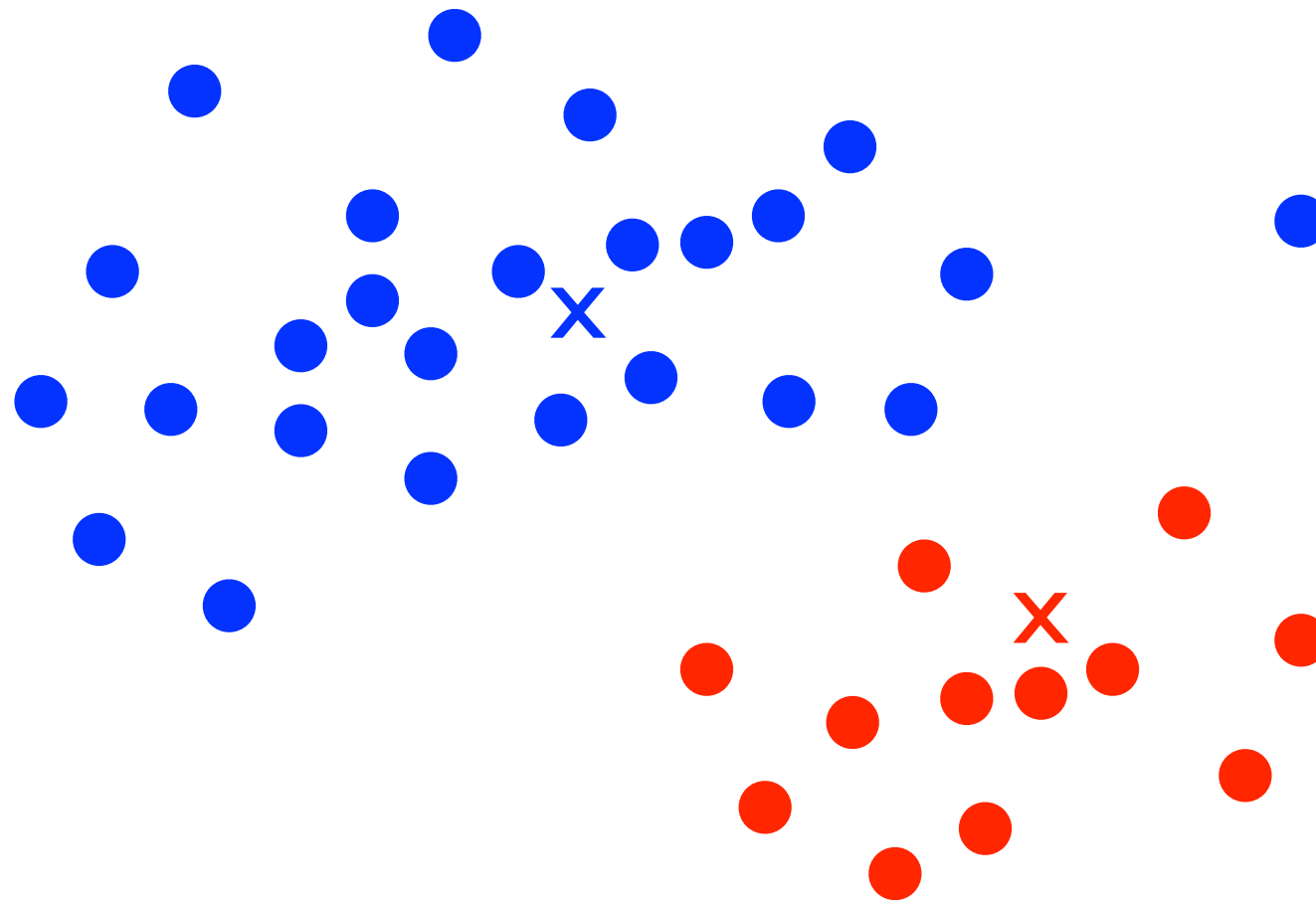
# K-means Clustering

- Step 4: re-assign each document to its nearest centroid



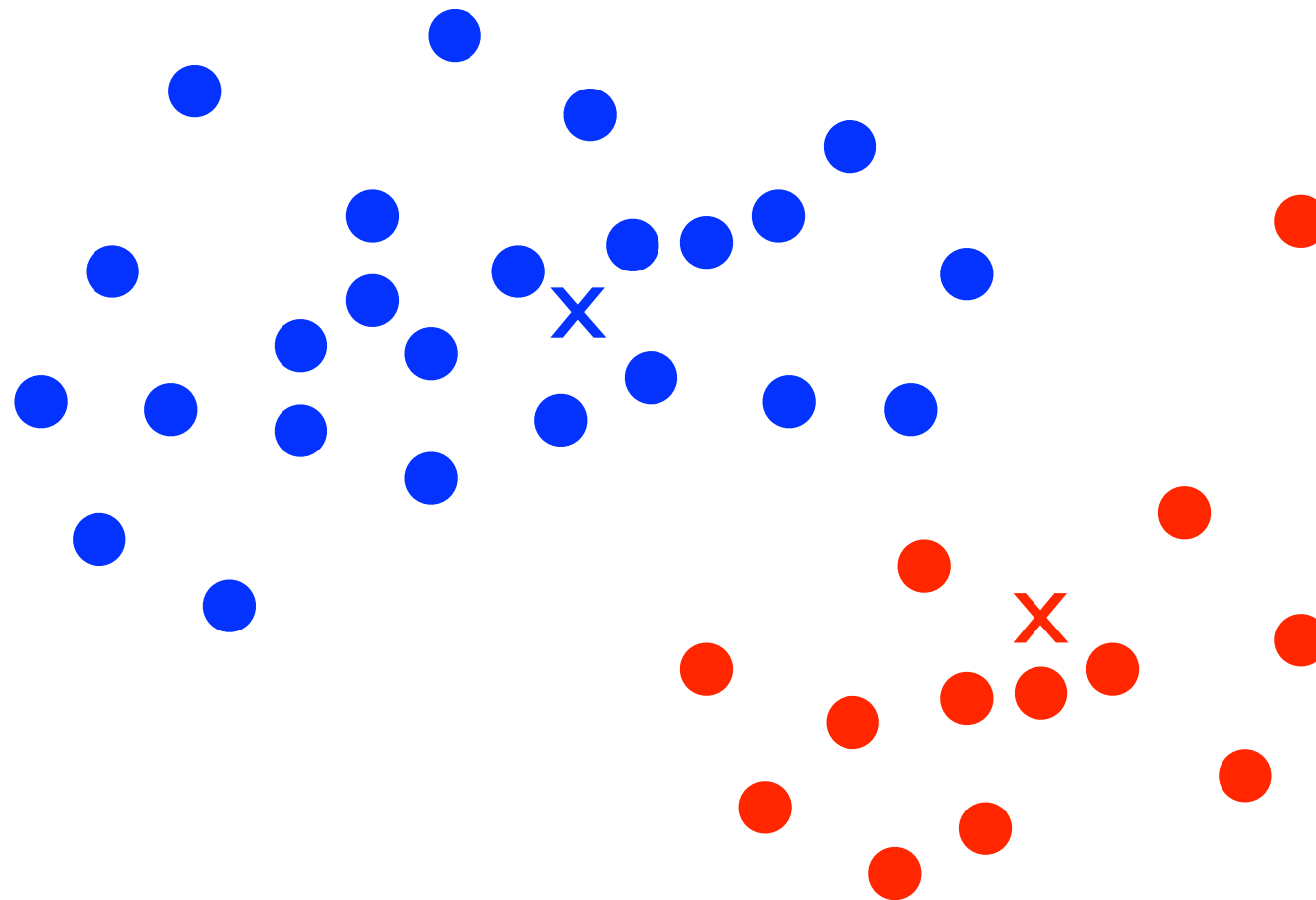
# K-means Clustering

- Step 4: re-compute all K cluster centroids



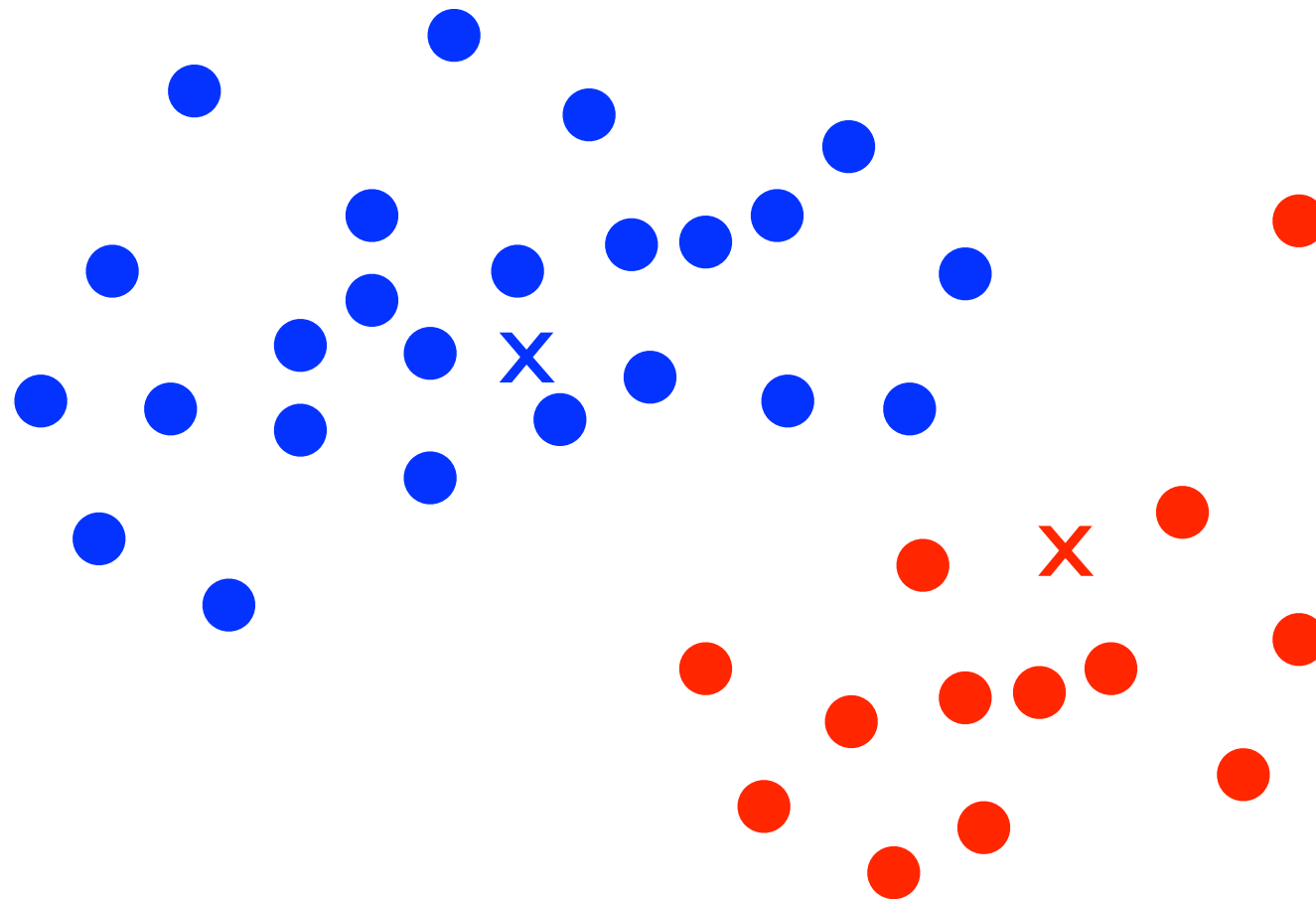
# K-means Clustering

- Step 5: re-assign each document to its nearest centroid



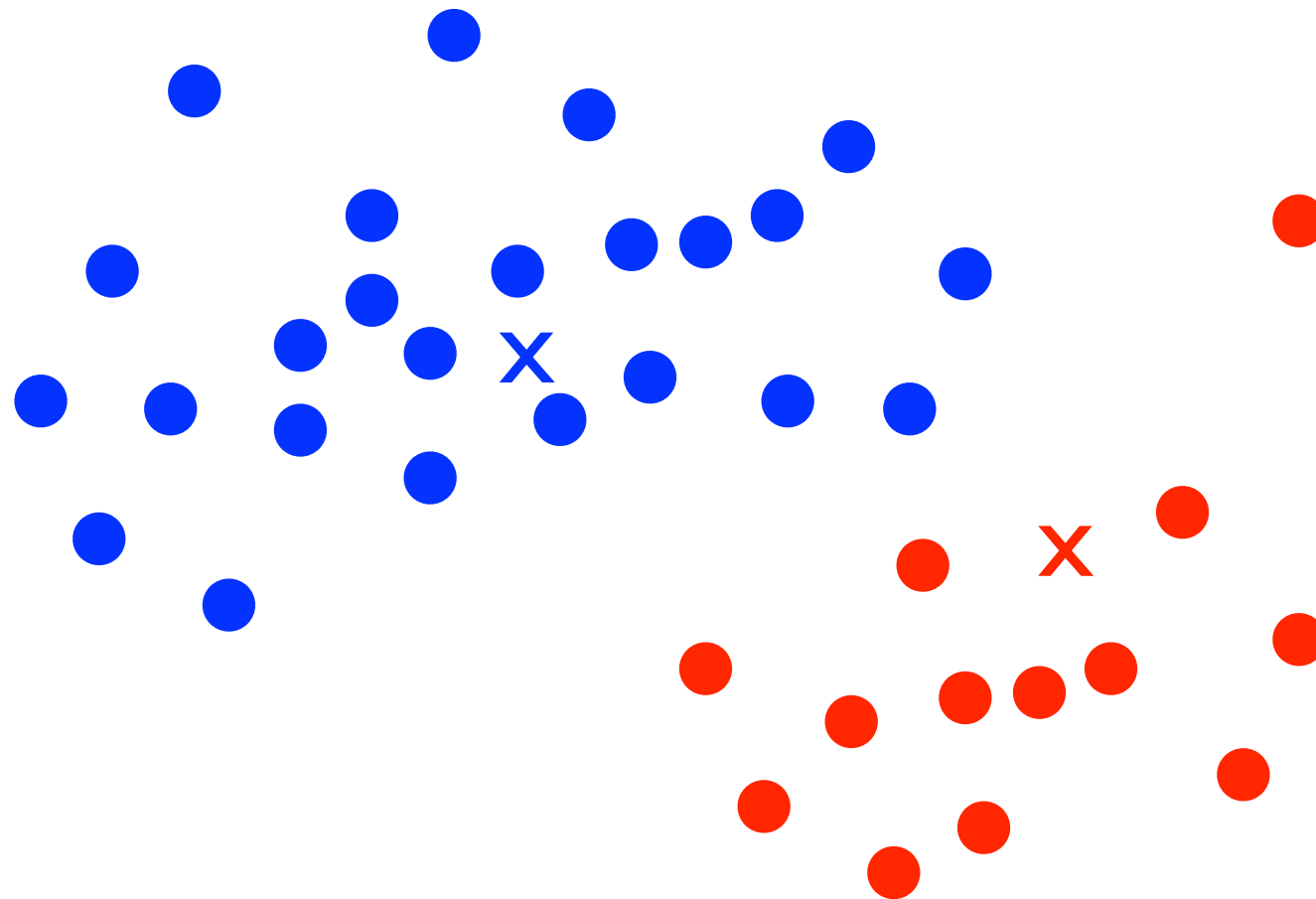
# K-means Clustering

- Step 4: re-compute all K cluster centroids



# K-means Clustering

- Step 5: re-assign each document to its nearest centroid



# K-means Clustering

- **Input:** number of desired clusters  $K$
- **Output:** assignment of documents to  $K$  clusters
- **Algorithm:**
  - ▶ **Step 1:** randomly select  $K$  documents (seeds)
  - ▶ **Step 2:** assign each document to its nearest seed
  - ▶ **Step 3:** compute all  $K$  cluster centroids
  - ▶ **Step 4:** re-assign each document to its nearest centroid
  - ▶ **Step 5:** re-compute all  $K$  cluster centroids
  - ▶ **Step 6:** repeat steps 4 and 5 until terminating condition



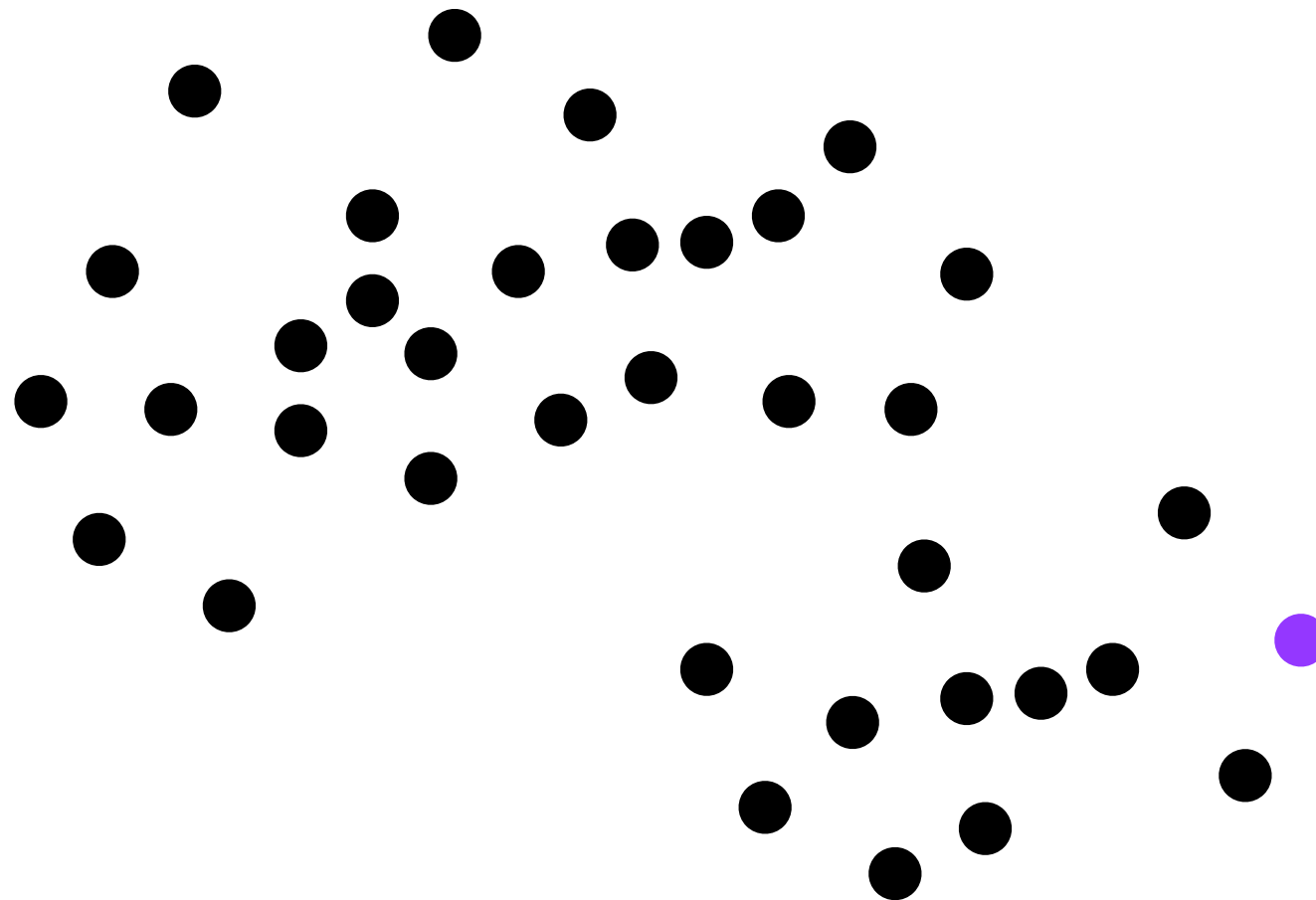
# K-means Clustering

## potential drawback

- The quality of the output clustering depends on the choice of **K** and on the **initial seeds**
- In many cases, the choice of **K** is pre-determined by the application
  - ▶ **Search engine results clustering:** grouping search engine results by topic
  - ▶ **Collection clustering:** grouping documents by topic to support navigation and exploration
- Later we'll see ways of setting **K** dynamically

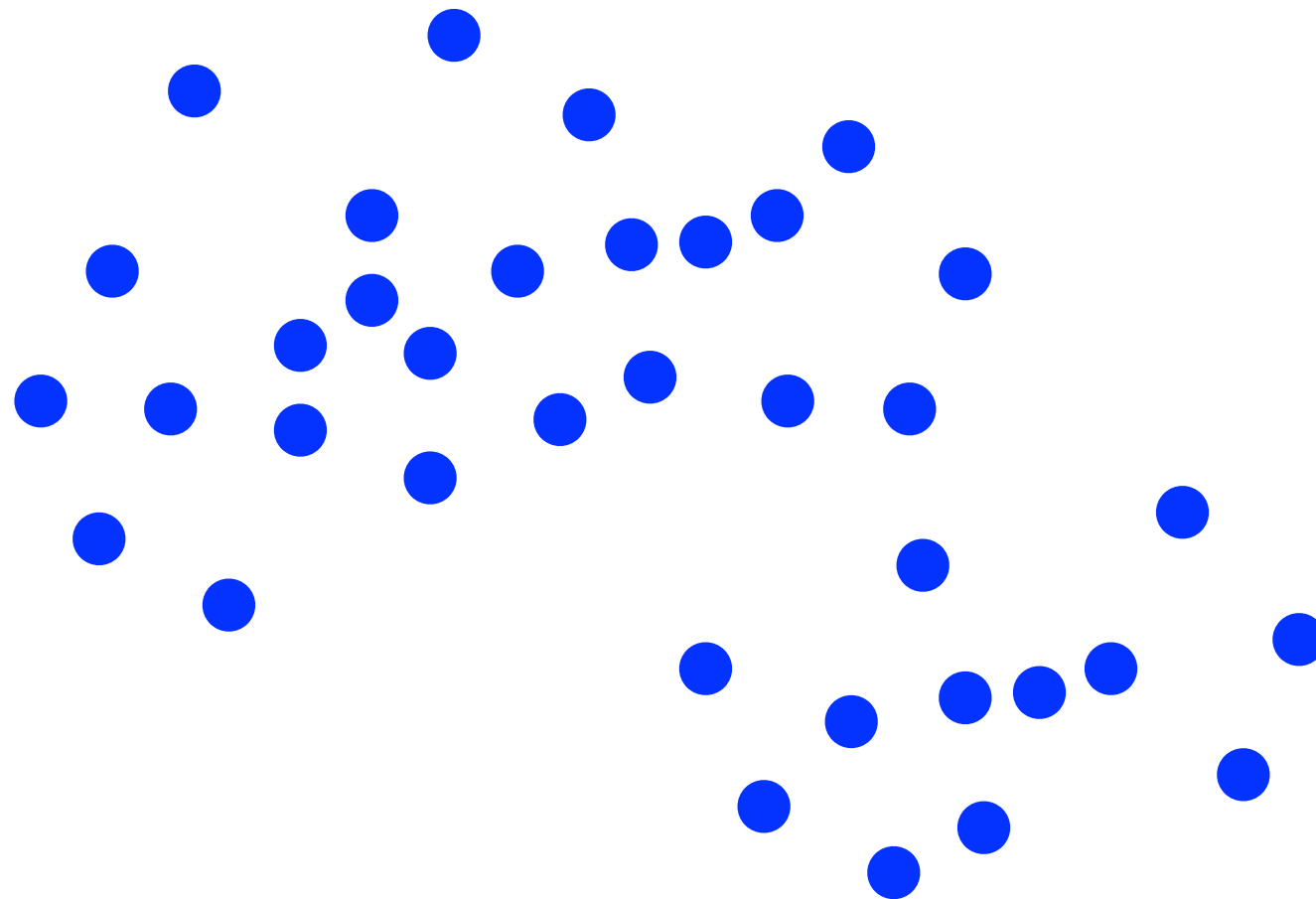
# K-means Clustering

bad seeds?



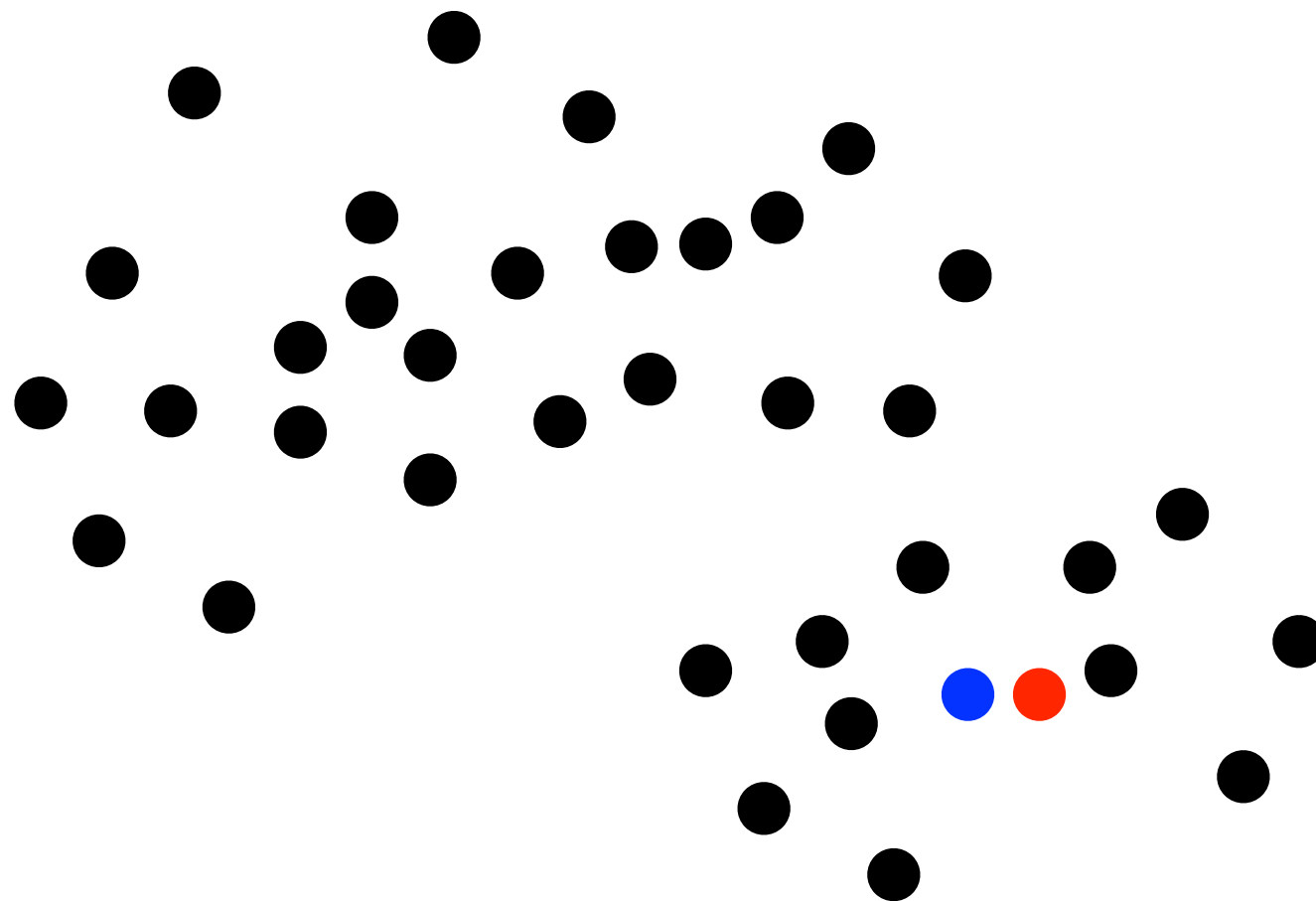
# K-means Clustering

bad seeds?



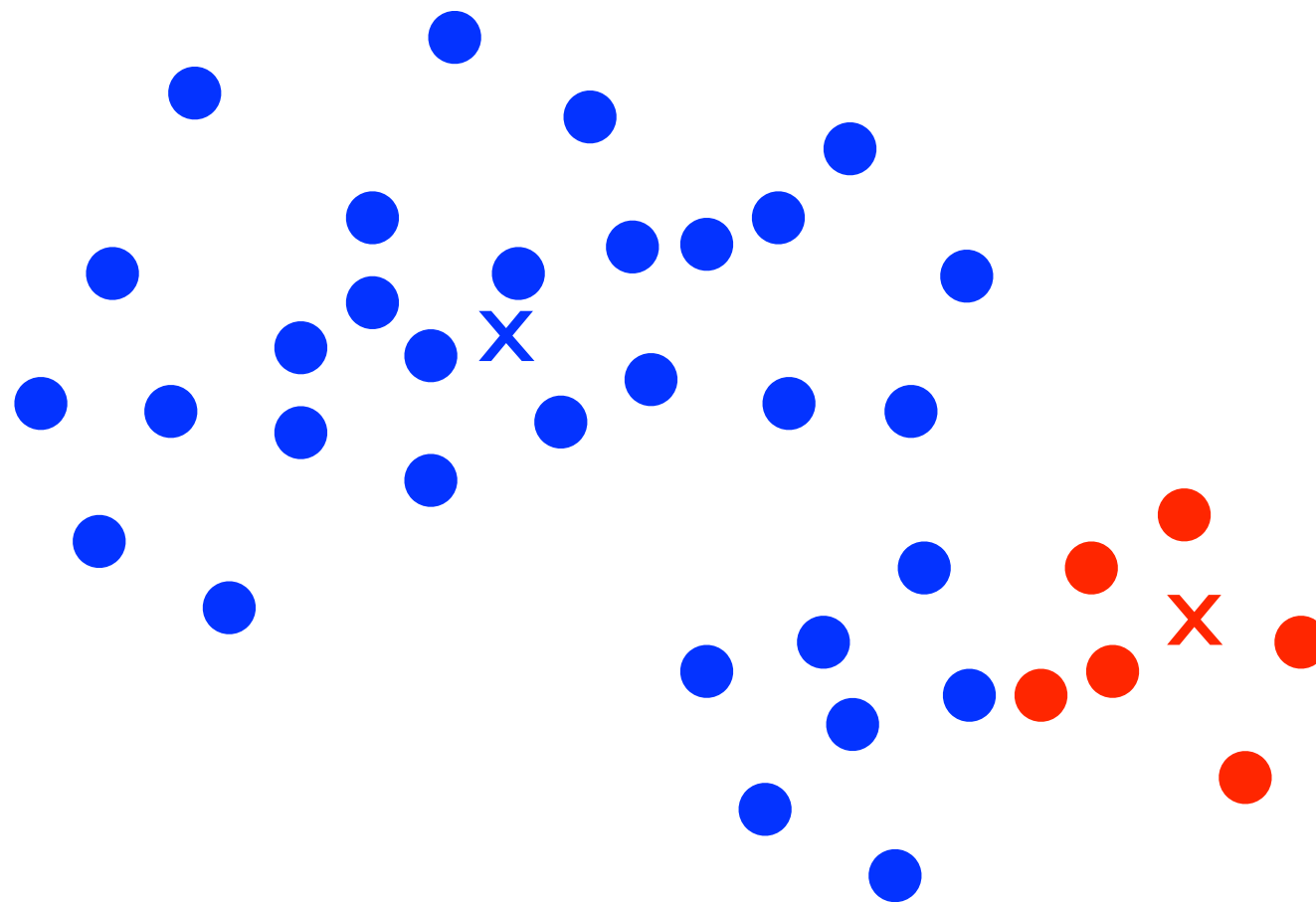
# K-means Clustering

bad seeds?



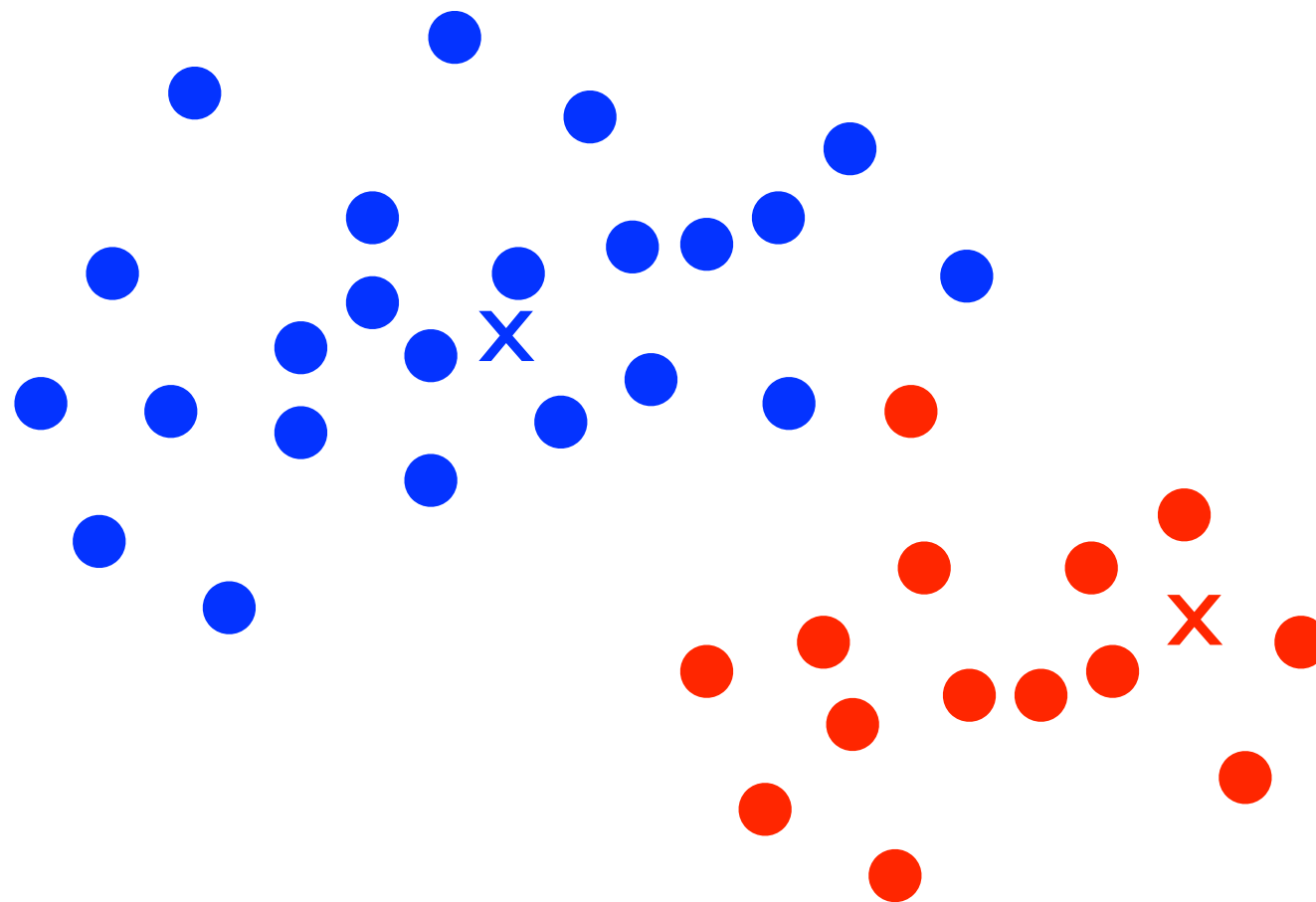
# K-means Clustering

bad seeds?



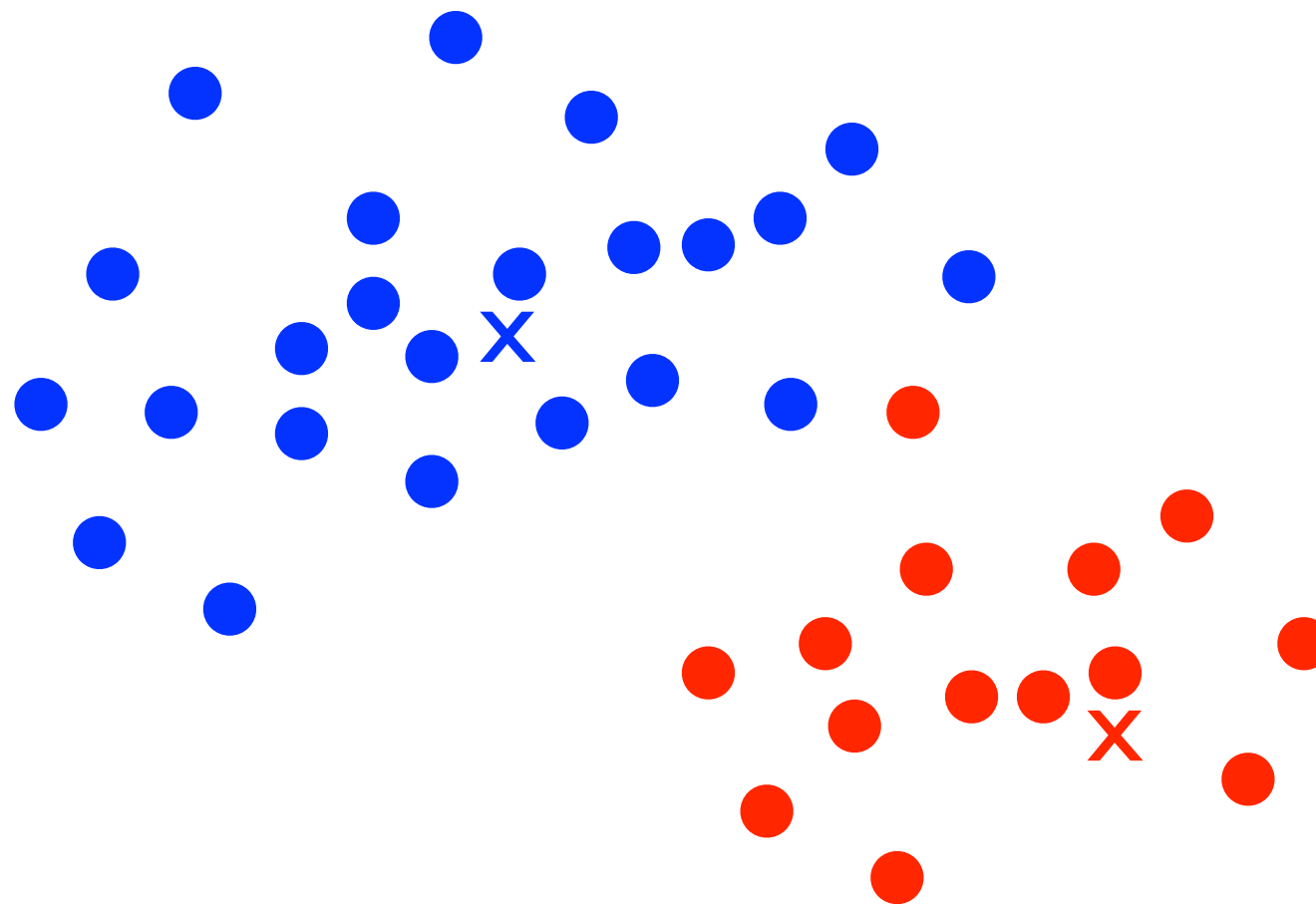
# K-means Clustering

bad seeds?



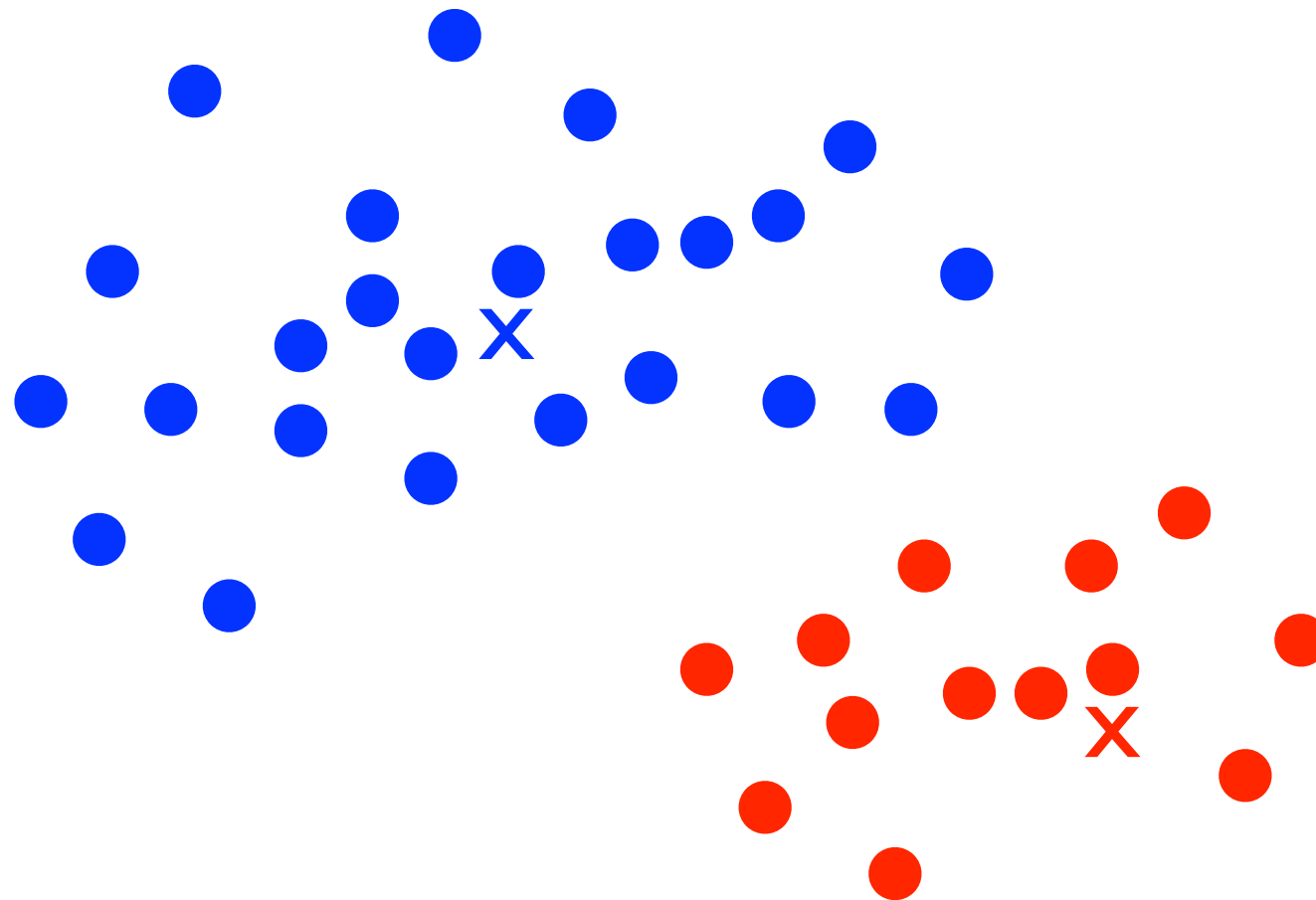
# K-means Clustering

bad seeds?



# K-means Clustering

bad seeds?





# K-means Clustering

## bad seeds

- It's difficult to know which seeds will yield a high-quality clustering
- However, it's usually a good idea to avoid seeds that are outliers
- How would you detect outliers?

# K-means Clustering

## clustering evaluation

- What does it mean for a clustering to be high quality anyway?
- What is the goal of clustering again?

# K-means Clustering

## internal evaluation

- In theory, a good clustering should have:
  - ▶ Similar documents in the same clusters
  - ▶ Different documents in different clusters

# K-means Clustering

## internal evaluation

$$\text{Clustering Quality} = \left( \begin{array}{c} \text{Average distance} \\ \text{between all pairs} \\ \text{of documents in} \\ \text{different clusters} \end{array} \right) - \left( \begin{array}{c} \text{Average distance} \\ \text{between all pairs of} \\ \text{documents in the} \\ \text{same cluster} \end{array} \right)$$

# K-means Clustering

## improved k-means

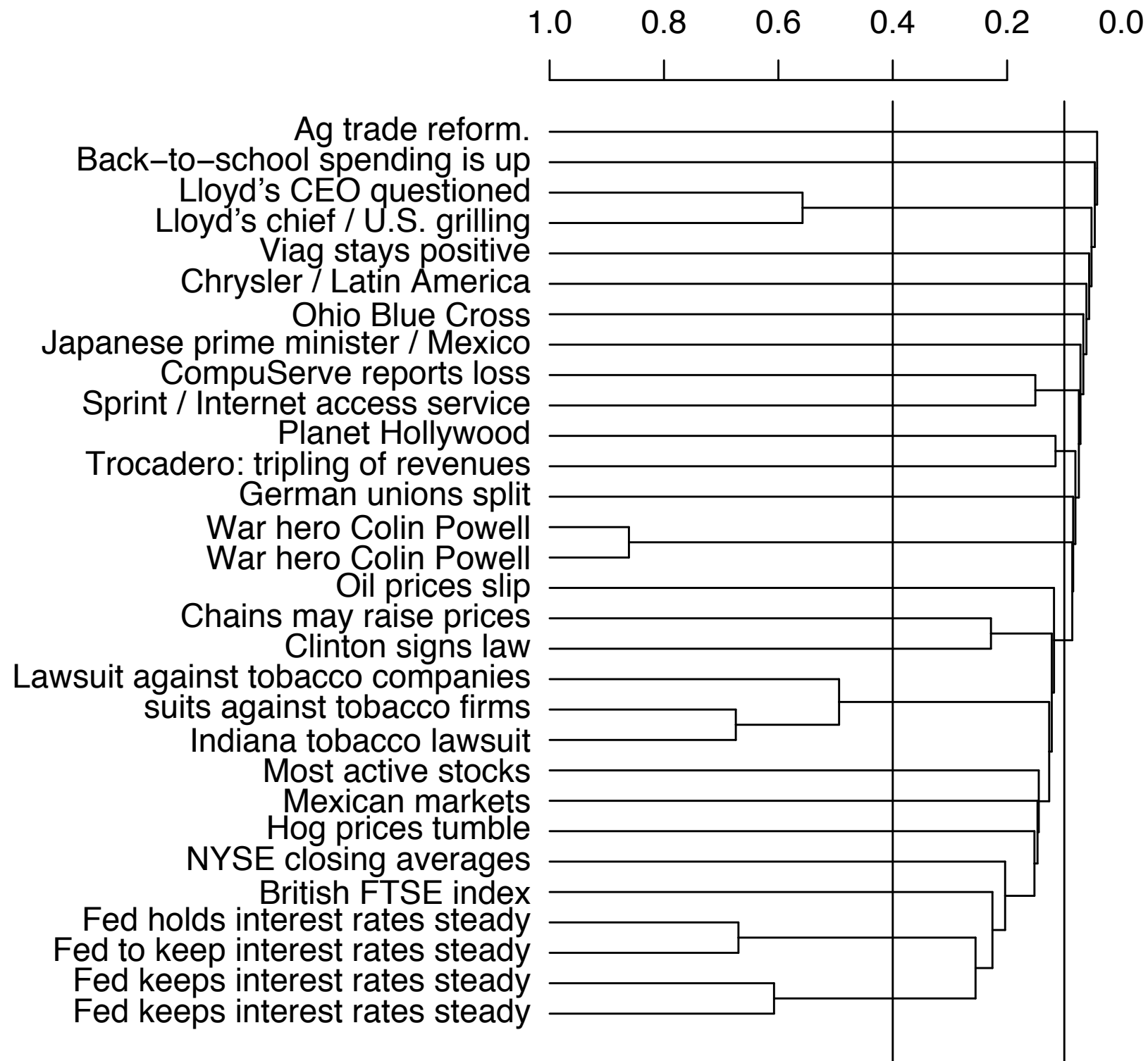
- Given a set of documents and a value  $K$ , run K-means clustering  $N$  times and keep the clustering that produces the greatest difference between the inter-cluster distance and the intra-cluster distance

# Bottom-up Agglomerative Clustering

# Bottom-up Clustering

- While K-means requires setting  $K$ , bottom-up clustering groups the data in a hierarchical fashion
- We can then set  $K$  after the clustering is done or use a distance threshold to set  $K$  dynamically (more on this later)

# Bottom-up Clustering





# Bottom-up Clustering

- Input: data
- Output: cluster hierarchy
- Algorithm:
  - ▶ Step 1: consider every document its own cluster
  - ▶ Step 2: compute the distance between all cluster pairs
  - ▶ Step 3: merge/combine the nearest two clusters into one
  - ▶ Step 4: repeat steps 2 and 3 until every document is in one cluster

# Bottom-up Clustering

- Input: data
- Output: cluster hierarchy
- Algorithm:
  - ▶ Step 1: consider every document its own cluster
  - ▶ Step 2: compute the distance between all cluster pairs
  - ▶ Step 3: merge/combine the nearest two clusters into one
  - ▶ Step 4: repeat steps 2 and 3 until every document is in one cluster

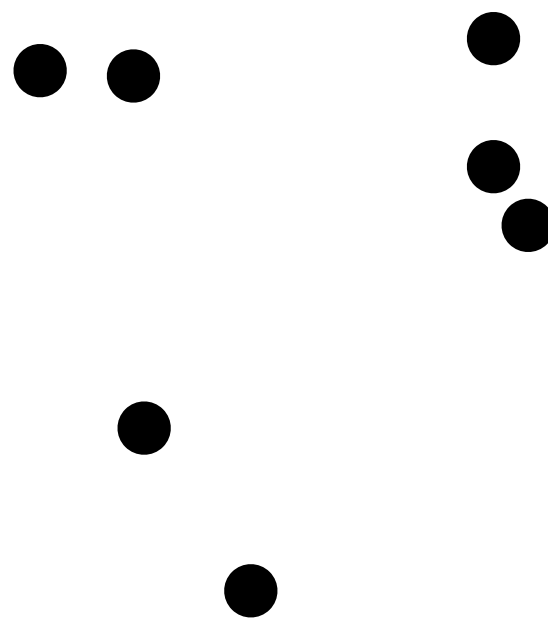
# Bottom-up Clustering

- Computing the distance between two clusters
- **Single-Link:** the distance between the two nearest documents
- **Complete-Link:** the distance between the two documents that are farthest apart
- **Average-Link:** the average distance between all document pairs (excluding those in the same cluster)
  - ▶ this is equivalent to using the distance between the two cluster centroids

# Bottom-up Clustering

## single-link

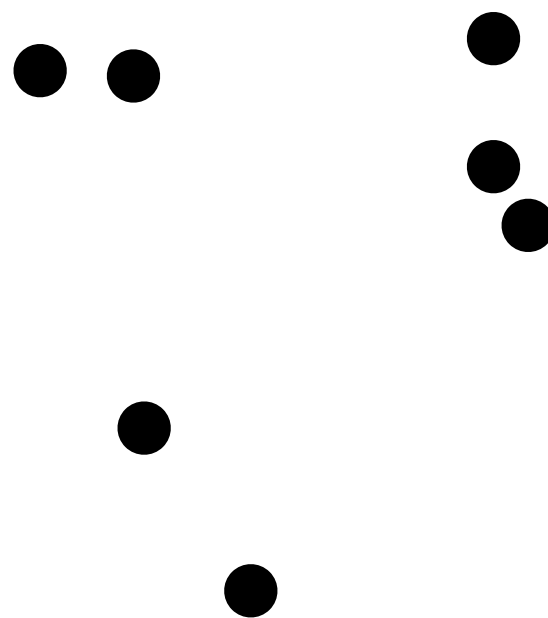
- Step 1: consider each document its own cluster



# Bottom-up Clustering

## single-link

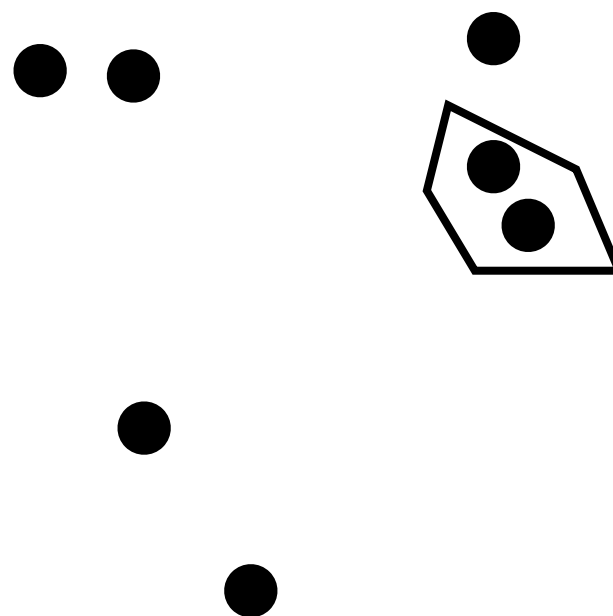
- Step 2: compute the distance between all cluster pairs
- Step 3: merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

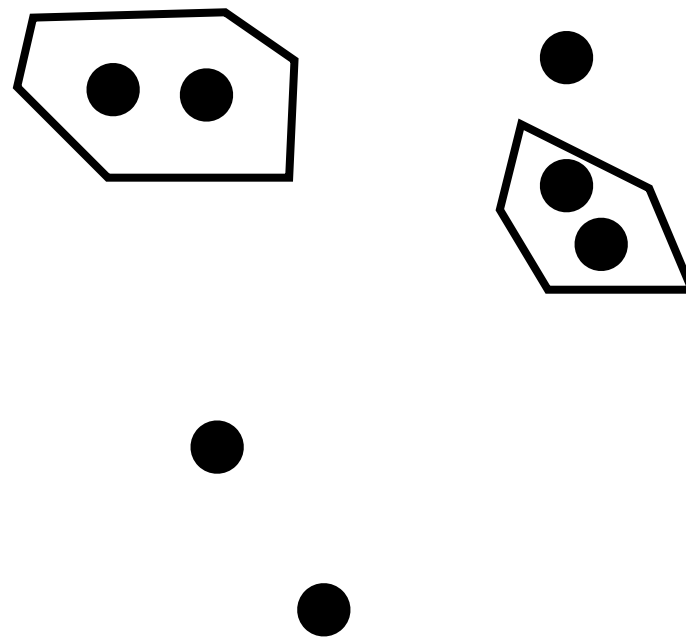
- Step 2: compute the distance between all cluster pairs
- Step 3: merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

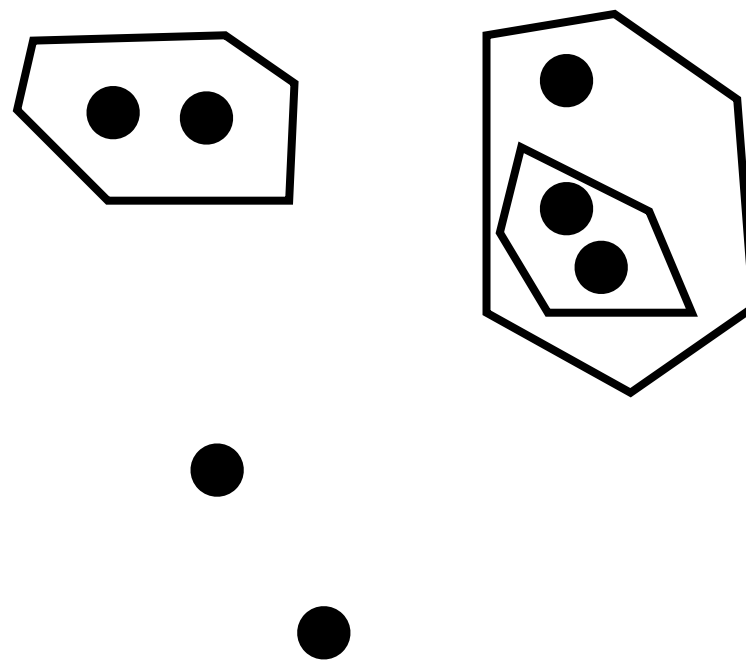
- Step 2: compute the distance between all cluster pairs
- Step 3: merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one

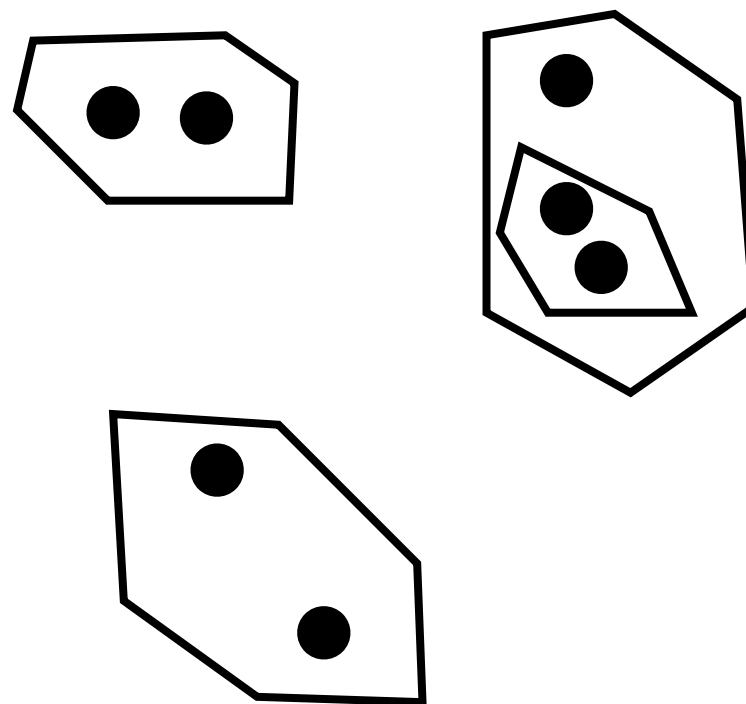




# Bottom-up Clustering

## single-link

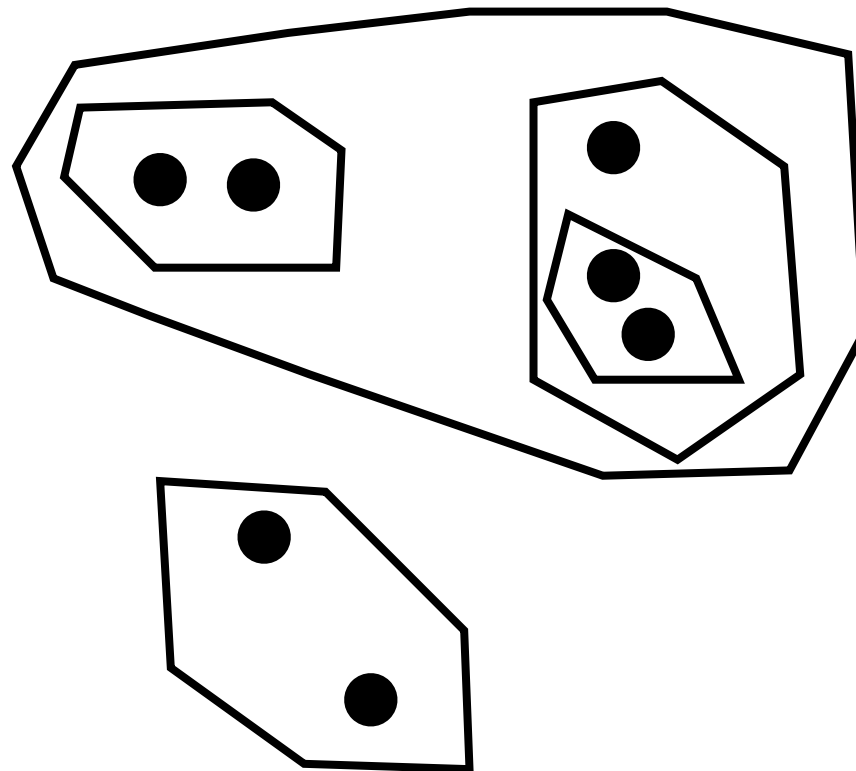
- Step 2: compute the distance between all cluster pairs
- Step 3: merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

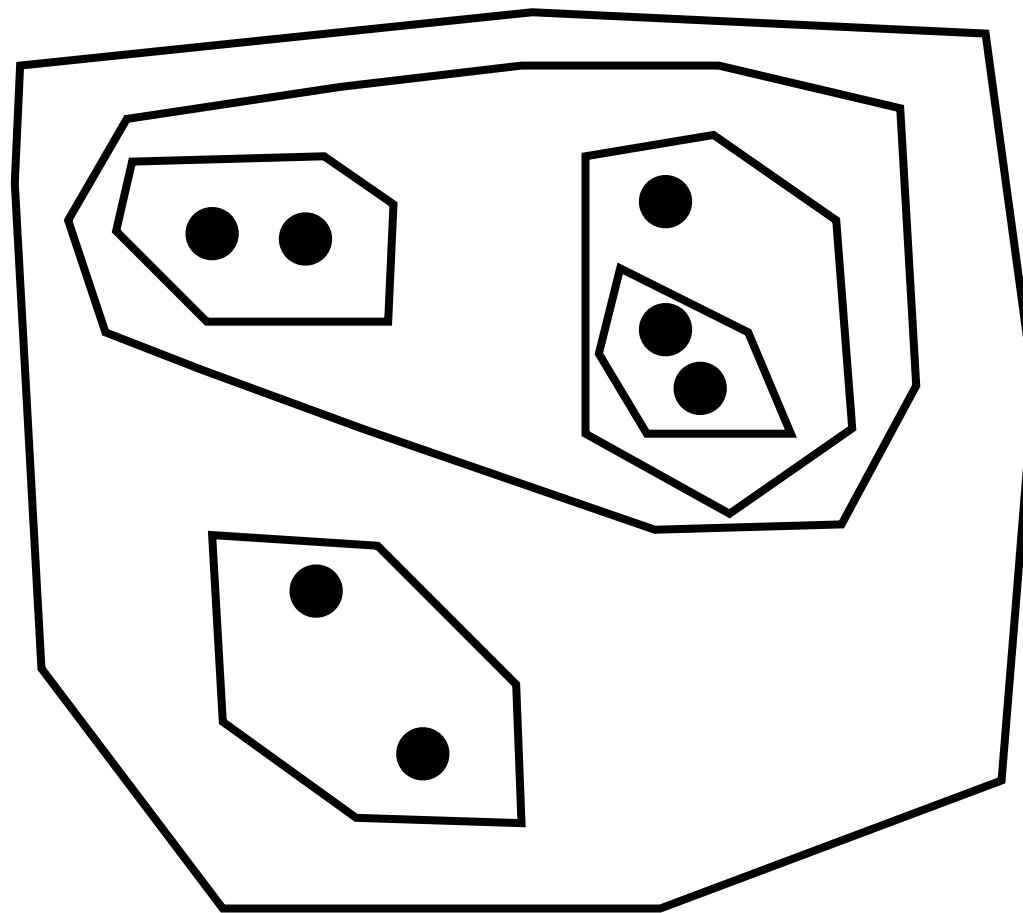
- Step 2: compute the distance between all cluster pairs
- Step 3: merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

- Step 2: compute the distance between all cluster pairs
- Step 3: merge/combine the nearest two clusters into one



# Bottom-up Clustering

- Setting  $K$  dynamically
- Instead of setting  $K$ , we could set a distance threshold  $T$
- Stop merging/combining clusters when the distance between the two nearest clusters  $> T$
- Using a distance threshold can help prevent “concept drift” (especially with single-link clustering)
  - ▶ text mining --> inls 613 --> unc --> basketball

# Bottom-up Clustering

## examples?

- Computing the distance between two clusters
- **Single-Link:** the distance between the two nearest documents
- **Complete-Link:** the distance between the two documents that are farthest apart
- **Average-Link:** the average distance between all document pairs (excluding those in the same cluster)
  - ▶ this is equivalent to using the distance between the two cluster centroids

# Labeling Clusters

# Clustering Applications

## collection clustering

The image is a screenshot of the Google News homepage. At the top is the Google logo and a search bar. Below the logo, the word "News" is displayed in red. To the right of "News" are two dropdown menus: "U.S. edition" and "Modern". On the left side, there is a vertical list of categories: "Top Stories", "World", "U.S.", "Business", "Elections", "Technology", "Entertainment", "Sports", "Science", "Health", and "Spotlight". The "Top Stories" category is highlighted with a blue border. Inside this border, a list of topics is shown: "Mitt Romney", "Chromebook", "Washington Redskins", "Earthquake", "Fidel Castro", "Cleveland Browns", "George McGovern", "Toronto Blue Jays", "Brad Pitt", "Jay-Z", and "North Carolina". A large black-bordered text box is overlaid on the right side of the page, containing the text: "How can we name clusters to inform someone about the kind of information they contain?". Below the text box, a news article is visible with the headline "Breaking: The area is on lockdown as authorities in Brookfield work to secure the scene, which is across the street from a shopping mall in Wisconsin." and a sub-headline "At Least 7 Injured in Spa Shooting". Below the article are several video thumbnails from various sources like The Associated Press, YouTube, CNN, CBS News, and Newsday. At the bottom, another article is partially visible with the headline "Romney, Obama in Dead Heat" from the Wall Street Journal.

Google

News

U.S. edition Modern

Top Stories

Mitt Romney  
Chromebook  
Washington Redskins  
Earthquake  
Fidel Castro  
Cleveland Browns  
George McGovern  
Toronto Blue Jays  
Brad Pitt  
Jay-Z  
North Carolina

World  
U.S.  
Business  
Elections  
Technology  
Entertainment  
Sports  
Science  
Health  
Spotlight

How can we name clusters to inform someone about the kind of information they contain?

Breaking: The area is on lockdown as authorities in Brookfield work to secure the scene, which is across the street from a shopping mall in Wisconsin.  
At Least 7 Injured in Spa Shooting

Romney, Obama in Dead Heat  
Wall Street Journal - 22 minutes ago  
By NEIL KING JR. Mitt Romney has strengthened his image as the candidate best able to boost the economy and has fought President Barack Obama to a near-draw on who can best serve as commander in chief, helping turn the 2012 election into a tie ...

# Labeling Clusters

## A simple solution

- Construct a vocabulary of terms and/or phrases (n-grams) that are frequent in the data
- Assign each cluster the term(s) or phrase(s) with the highest mutual information



# Mutual Information

$$\text{MI}(w, c) = \log \left( \frac{P(w, c)}{P(w)P(c)} \right)$$

- **P(w,c)**: the probability that a document contains word **w** and belongs to cluster **c**
- **P(w)**: the probability that word **w** occurs in a document from any cluster
- **P(c)**: the probability that a document belongs to cluster **c**

# Mutual Information

$$\text{MI}(w, c) = \log \left( \frac{P(w, c)}{P(w)P(c)} \right)$$

- If  $P(w, c) = P(w) P(c)$ , it means that the word  $w$  is independent of cluster  $c$
- If  $P(w, c) > P(w) P(c)$ , it means that the word  $w$  is not independent of of cluster  $c$

# Mutual Information

- Every document falls under one of these quadrants

total # of instances  $N =$   
 $a + b + c + d$

$$P(w, c) = a / N$$

$$P(c) = (a + c) / N$$

$$P(w) = (a + b) / N$$

$$MI(w, c) = \log \left( \frac{P(w, c)}{P(w)P(c)} \right)$$

	belongs to cluster <b>c</b>	does not belong to cluster <b>c</b>
contains word <b>w</b>	a	b
does not contains word <b>w</b>	c	d

# Summary

- Clustering: grouping similar documents (or instances) into subsets
- Exploratory analysis: the goal is to discover common and uncommon properties of the data
- K-means and Agglomerative Bottom-up Clustering (there are many, many others)
- Labeling clusters